11

Patterns of Distribution

		Page
Introduction		11-1
11A.	Some matters of principle	11-2
11B.	Variations on the stem-and-leaf theme	11-5
11C.	Variations on the histogram	11-7
11D.	Intervals of different lengths	11-17
11E.	How to picture and compare distributions when the responses are ordered, not numbered	11-24
11F.	How to deal with distributions of counts -- especially when J-shaped	11-28
11G.	How far have we come?	11-40

*****

11R.	Xmas-tree patterns	11-47

*****

11Y.	Problems	11-51

Location of tables and plots for 11

Number	Page
1	11-6
2	11-8
3	11-9
4	11-11
5	11-12
6	11-14
7	11-16
8	11-19
9	11-20
10	11-22
11	11-23
12	11-25
13	11-27
14	11-29
15	11-30
16	11-33
17	11-34
18	11-36
19	11-38
20	11-41
21	11-42
22	11-43
23	11-45
24	11-48
25	11-49
26	11-50

11

Patterns of Distribution

How are a set of observed numbers distributed? Any answer to this question, however sophisticated its mechanism, is likely to begin with counts -- counts of how many cases fall in each of certain intervals or counts of how many cases take on each of the possible values. The devices that we use to display apparent behavior -- or to compare observed behavior to a fitted partial description -- need to be devices appropriate to counted data. Now that we have a little background in dealing with counts, we are ready to tackle the question of exhibiting how counts are distributed among classes -- of what counted distributions look like -- and how our looks at them can be made more searching and more effective. We will need a new principle or two, added to those that are now becoming familiar, but we will find that looking at counted distributions again reveals many of the same aspects and approaches as have our other topics.

In doing this we have to recognize the differences among different kinds of classifications:

- classes that have names but are not ordered according to some definition of $<$.

The countries of Europe or the countries of sub-Saharan Africa would be examples of this.

- classes that are arranged in order according to some $<$.

These may be classes that exist in themselves or classes defined for our convenience. Justices of the Supreme Court rank above Senators, who rank above Congressmen, the boundaries being widely recognized and sharp. Colonels, Majors, Captains and Lieutenants would provide another example of classes with independent existence. Classes of people aged 20-24, 25-29, 30-34, 35-39, and so on offer a simple example of classes defined for our convenience.

In most instances where ordered classes exist for our convenience, there is an underlying numerical scale. This case will be the focus of our interest in this chapter.

11A. Some matters of principle

We already have some background that is unfamiliar to many; we have been using stem-and-leaf arrays to display distributional information. As one particular result, we are used to seeing pictures of patterns of values in which the different values are spread out vertically.

It cannot be a matter of too high principle whether pictures of distributions spread the values out horizontally or vertically. Clearly both can serve us. But it is important for us to understand why sometimes one choice is made and sometimes another.

Whether the usual division between what is vertical and what is horizontal in a graph is intrinsic in *Homo sapiens* or is an accident of Western civilization is a difficult question, one that is hard to even see how to attack.

Fortunately, we do not need to know its answer. We live in a world in which responses are plotted up and down, while quantities that describe alternatives -- whether or not it is likely that they "affect" the responses -- are plotted from left to right. (Is the fact that North is at the top of our maps a consequence of the West-East trend of the Mediterranean Sea?)

There are two quite disparate views about a plot of a distribution of numerical values. One view thinks -- often only implicitly -- about the numbers of cases (per interval or per possible value) as what matters. It usually thinks of these numbers as responses and about the values involved as something that influences or describes the environment of such counts. Those who -- for the moment at least -- think in this way naturally spread values out horizontally and plot counts vertically. (Within our tradition we could expect nothing else.)

The other approach, typified by the stem-and-leaf array, thinks -- again often only implicitly -- of the values as what matter -- as what ought to be displayed clearly. Those of this mind will spread values out vertically, and use some graphical device to make clear the frequency with which various ranges of values occur.

Each approach has its place. We should all try to use each in its place.

How do we picture importance most directly? How could we convey relative frequencies visually to a preliterate --

be he uneducated Australian bushman or preschool child? We leave the final answer to this question to the psychologists, but give a simple approximate answer -- pebbles.

To show one thing 100 times as important -- or 100 times as frequent -- as another we could use 100 pebbles for one but one pebble for the other. To make clear the comparison we would be very likely to lay out the 100 pebbles side by side -- but probably not in a line. A square 10x10 array of pebbles will make the point, and be easy to comprehend. (In the strict word-genealogy sense of "comprehend" -- grasping together.) All of us have played with pebbles and blocks and many similar things, whether we remember this or not. Almost all of us have developed at least somewhat of a feel for area -- for area as representing importance or count. As a result, whenever we can, without great loss of some other sort, use a scheme of display in which counts are proportional to areas, we ought to do so.

The problems of effective representation of counted distributions, then, revolve around how to combine as much as possible of:

- proper use of vertical or horizontal as the direction of spreading out values. (The cultural principle.)
- sticking to area proportional to count. (The area principle.)
- gaining graphical effectiveness by making the (implicit) standard of comparison a straight line,

preferably a horizontal (or vertical) one. (The comparison standard principle.)

- seeking out the advantages of making residuals evident and easily scrutinized. (The residuals principle.)

- following out that experience which favors expressing counts by their square roots. (The square-root principle.)

Two of these principles are new to us, one comes from chapter 10, the other two are old. To combine all five in a single picture often calls for real ingenuity.

11B. Variations on the stem-and-leaf theme

Stem-and-leaf displays meet our two new criteria: Values are spread out vertically. Using one digit per case means that the area covered with digits is proportional to count -- at least on a typewriter. Thus it will help us to prepare for more complex cases if we look at a few variations on the skeleton of the stem-and-leaf structure.

Display 4 of chapter 8 (page 8-7) showed 120 second residuals from a Nebraska voting example. It did this in what we now consider a conventional stem-and-leaf pattern. Table 1 repeats this data in the same basic stem-and-leaf pattern, swapping up with down so as to leave positive values higher than negative ones.

So long as we think of a stem-and-leaf display as something printed -- as a conventional table or something produced line by line by a computer printer -- it is reasonable for it to begin with the most negative values and move toward

table 1 of chapter 11

Plus-side-up Version of display 4 of chapter 8

(120 residuals from an elementary analysis of a two-way table)

+5	4
+4	22
+3	118 339 26
+2	702 431 448 633 842 0
+1	263 133 367 949 792 250 014 3
+0	181 413 624 267 768
-0	167 048 443 743 5
-1	266 128 306 576 616
-2	930 016 070 791 645 6
-3	452 212
-4	203
-5	17
-6	1

(Latest corrections to display 4 not incorporated.)

the most positive ones. If we are to think of a picture, however, the most positive values should be at the top.

The simplest way to give the same information purely graphically -- we must stress purely because any stem-and-leaf display is largely graphical -- is to use "bars flying from a mast", in which the length of each bar is proportional to the corresponding count. Plot 2 shows the result. Except for the elimination of some distraction, this has very little to offer beyond table 1. If we were dealing with large counts, where direct use of the stem-and-leaf display would be impossible or confusing, the bars would have a real advantage.

We ought next to see what we can do to give effect to last chapter's view that counts are most usually well expressed numerically by their square roots. At first sight, such a square-root principle seems squarely in contradiction with the area principle, which calls for equal areas for equal counts. But this need not be so. Bars have both height and length. If we consider bars both of whose dimensions increase as the count increases, while retaining the same proportion, we can have the linear dimensions of our bars proportional to the square root of the count, while their area is proportional to the count itself.

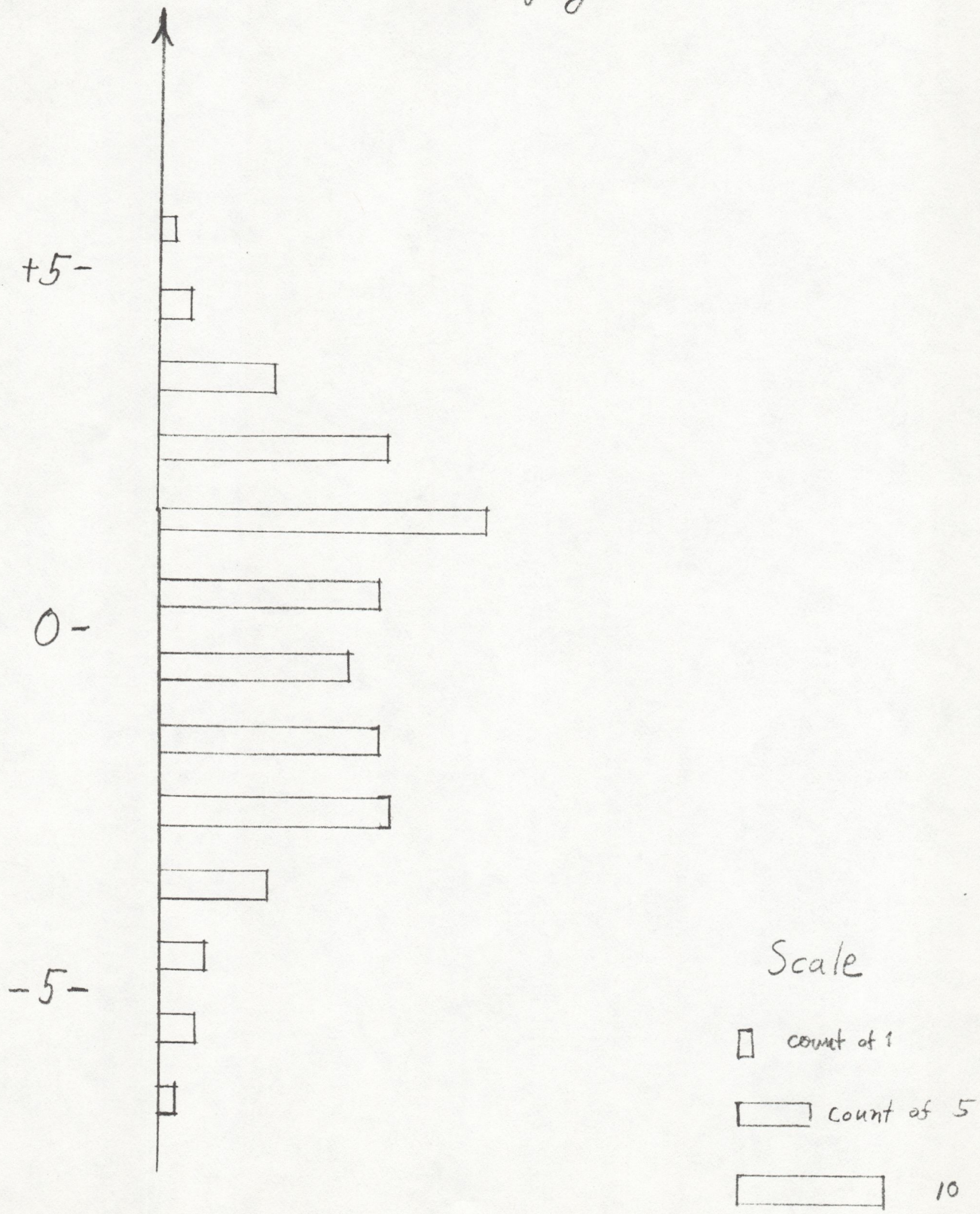
Plot 3 shows the same distribution of values described by such bars, bars which are naturally called (square) rooted bars, or even "rootobars" .

11C. Variations on the histogram

Our last picture admits two different interpretations -- either separately or together. According to the first, it is

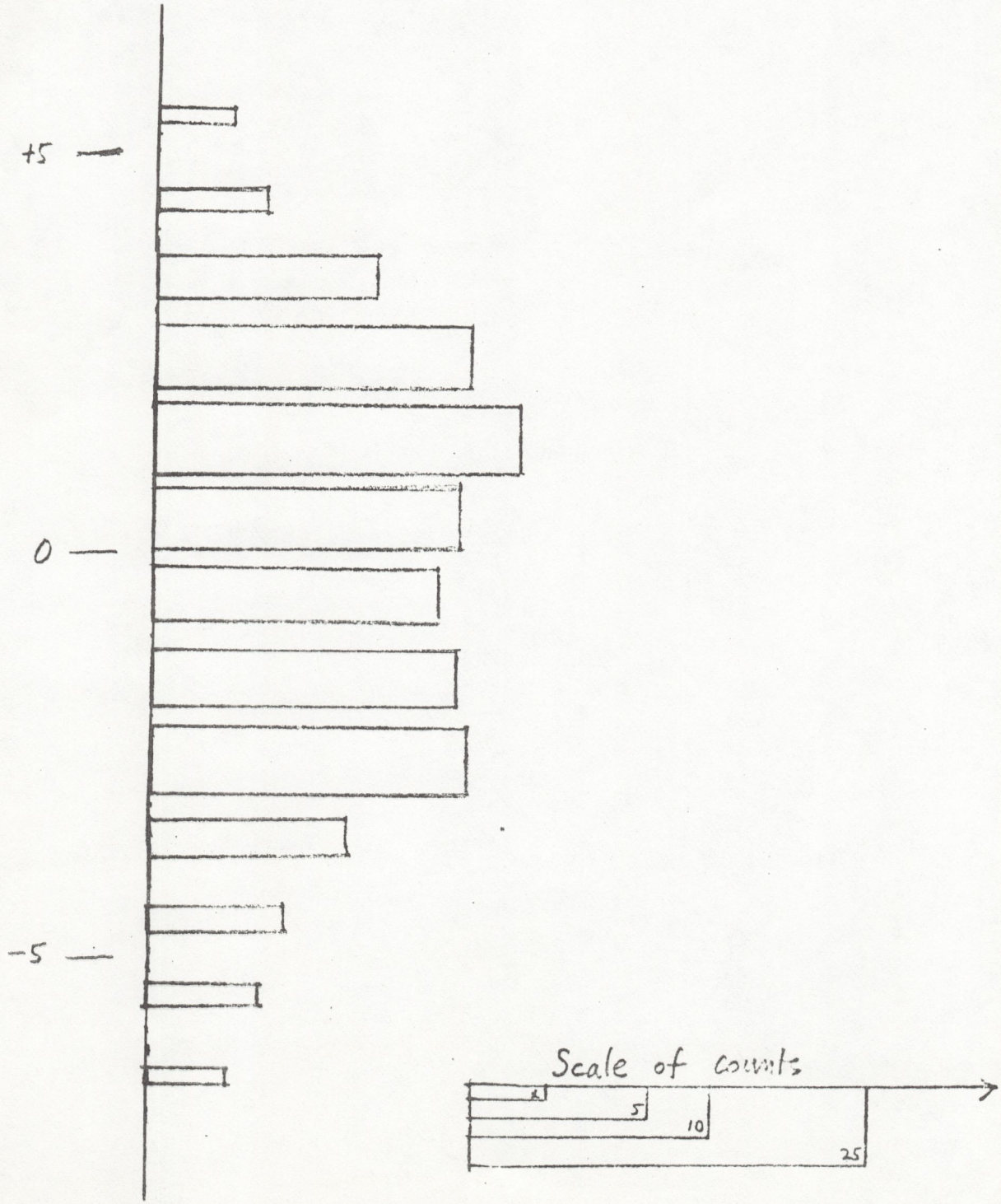
plot 2 of chapter 11

The 120 residuals shown by "bars flying from a mast"



plot 3 of chapter 11

The same 120 residuals, shown by histogram "lying from a wall".



a display of the distribution of the values of 120 residuals, with equal area given to each value. As such, it is rightly plotted with the values (of the residuals) running down to up.

According to the second interpretation, it shows -- for each interval of length one that can be defined by giving the units digit -- the count associated with that interval, this count being expressed on a square root scale. To the extent that this latter interpretation is correct, our picture would be more appropriately plotted with values running from left to right, thus reserving down to up for count information which is now clearly regarded as a response.

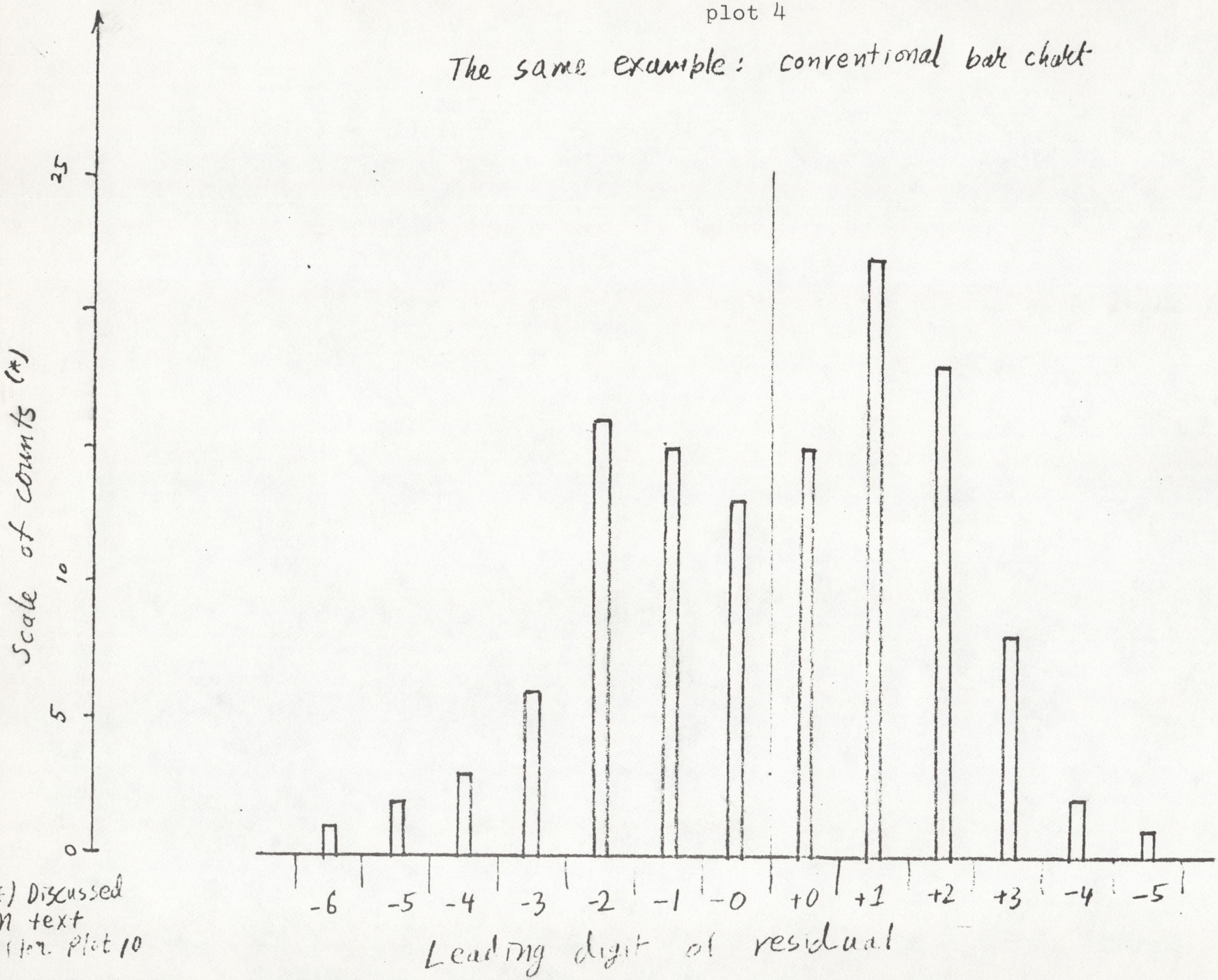
Let us, then, look at the most relevant forms of picture for showing the distribution of these 120 residuals with values running from left to right. The form most likely to appear in a newspaper is a bar chart. Plot 4 provides an example, about which little comment is needed.

For the professional -- who is used to thinking (at least abstractly) about distributions involving very many possible values and behaving unrealistically smoothly -- there is something wasteful in the spaces between the bars of a bar chart. He is likely to make use of a <u>histogram</u>, like plot 5, instead. (Especially since half as many vertical lines mean noticeably less drafting effort.)

No approach so far used in this section takes any account of the virtues of having a count expressed as a square root. (It was these virtues, we recall, that led to plot 3 -- and thus to our desire to try approaches with the values spread left to right.) If we do not feel attached to the principle

plot 4

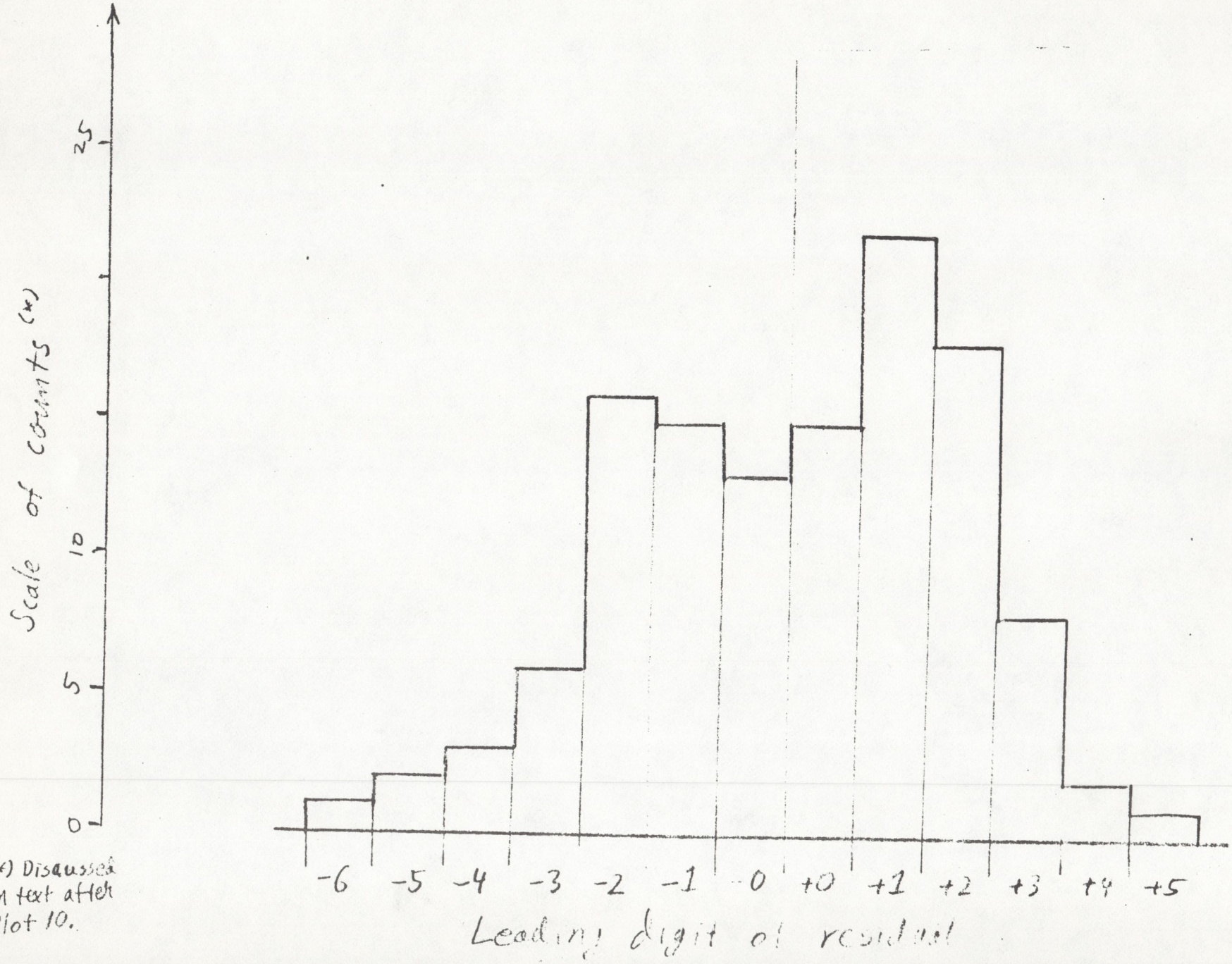
The same example: convention bar chart



(*) Discussed in text after Plot 10

plot 4 of chapter 11

The by-now-familiar example: histogram



(*) Discussed in text after Plot 10.

of equal area for equal count -- some of us may feel just this way -- we can bring in the square root most easily in a picture differing from a histogram in but one way: The heights of the rectangles are now made proportional to the square roots of the counts, rather than proportional to the counts themselves. Though probably bad philology, it makes memory easier to call such a plot a "rootogram" . Plot 6 shows our, by now familiar, 120 residuals as a rootogram.

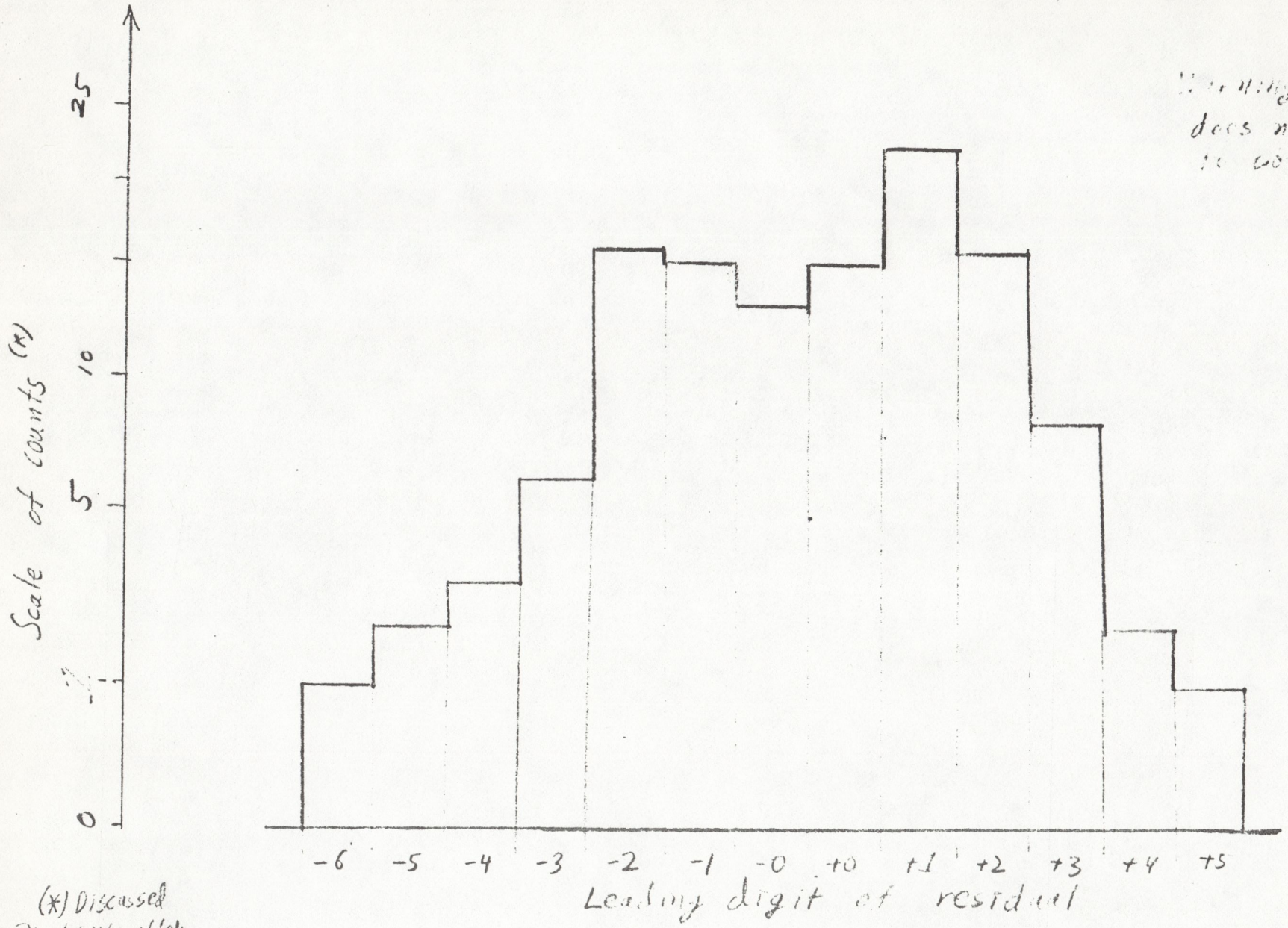
When we look at plot 6, especially when we compare it with plot 5, we see at least two things of major importance:

- the irregularities of outline now seem to be of more nearly equal size all across the distribution.
- the tails are now represented by larger, more noticeable blocks, so that we can see much more clearly what is -- or is not -- happening at or near them.

These observations seem to apply to the great majority of distributions of sets of values that arise in practice. They are part of the reason why square-root scales for counts are so useful in plotting distributions.

Some will wonder why we think it good that irregularities of outline have become more uniform in size. Is it not true, they will ask, that the search for revealing irregularity has been one of our main tasks? Are we not covering up just the sort of thing which we should be looking for? It is true that we seek revealing irregularity, but our problem is often -- indeed mainly -- a search for r e v e a l i n g irregularity against a background of unremitting, unrevealing, r o u t i n e irregularity.

plot 6 of chapter 11



(*) Discussed in text after Plot 10

Plot 6

The by-now-familiar example: histogram

What we tend to do by making the irregularities that occur commonly (that is, frequently and routinely and under circumstances where there is no indication from further analysis that we should have detected a clue or clues) of roughly constant size is to give the uncommon irregularity both a more uniform chance -- and a better chance -- of being detected. The routine causes of routine irregularity are many -- usually no one cause dominates, nor, indeed, does any one class of causes dominate. If we can make the collective effect of these routine causes roughly uniform in size, we almost always have a much better background against which to spot the kind of irregularity that will help us.

Coming back to our step-by-step advance, we can now, at the price of a little extra effort restore the area principle while keeping the vertical square root scale. We can make a "rootobar" plot according to the same principles and practices that we used in plot 3, merely interchanging horizontal and vertical axes. Plot 7 shows the result for our familiar 120 values.

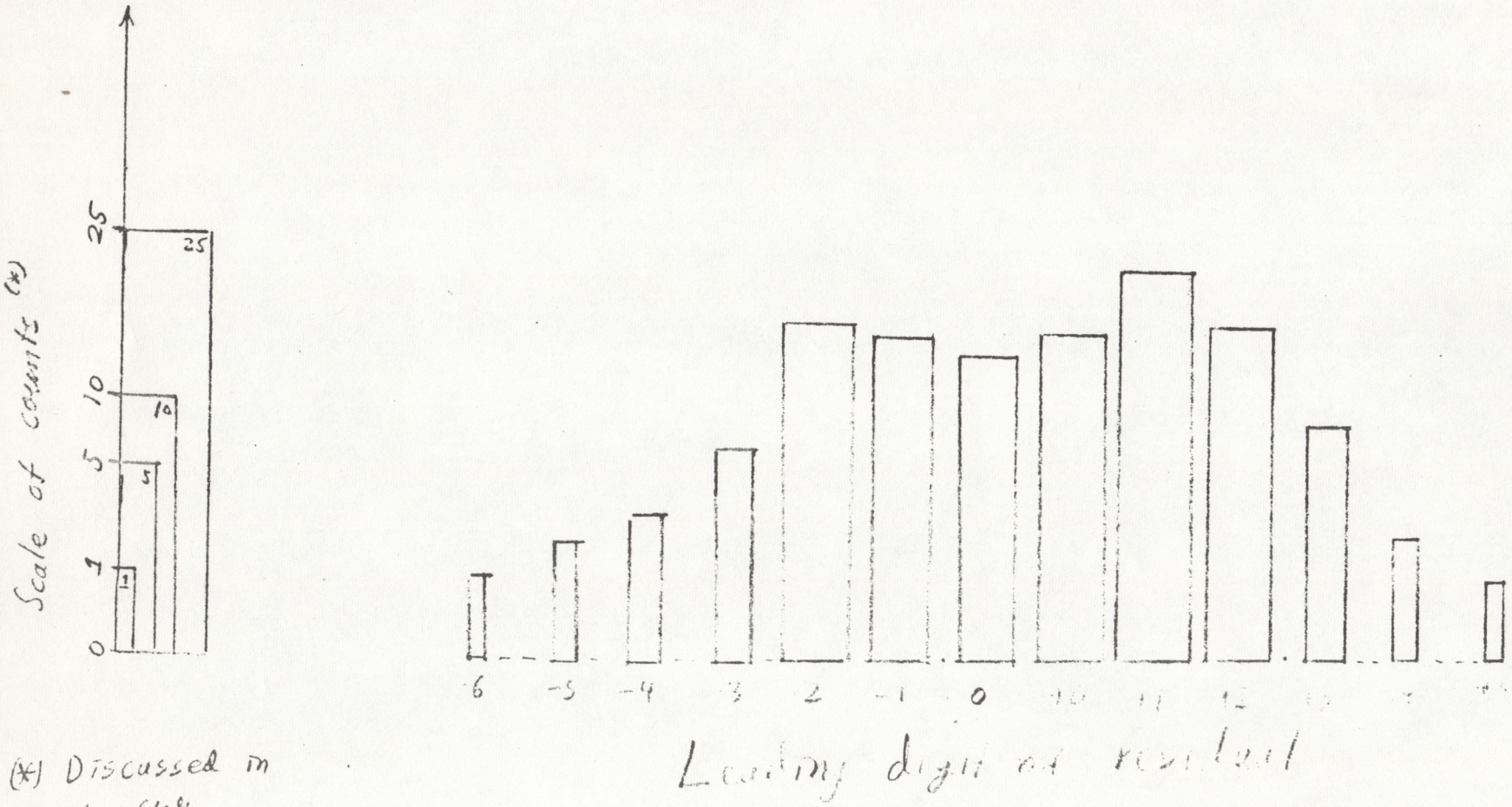
Comparing this plot with plot 6 we see that it has preserved almost all the advantages of a rootogram, while giving a more appropriate feeling of how the values thin out as we move toward either tail. (What matters here is exactly a "feeling". The essential question is: Does the viewer gain the correct emotional reaction to this picture?).

Different analysts will have different feelings about relative merits of different choices among the six forms of

plot 7 of chapter 11

Plot 7

The by-now-familiar example: conventional rootchart



(*) Discussed in text after Plot 10

patterns-of-counts plots we have shown in this section and the last (and, as well, among a few more variations that we have not shown). Indeed it is quite proper for any of us to have one order of preference when making one particular application and another for another -- or, more generally, others for others.

If the writer had to specify general rules fixing a single preference order for each general class of situation, he would choose to insist on a square-root scale above all else. If drafting effort was crucial, he would accept a rootogram (as in plot 6), and give up the area principle. If it was important not to shock rather naive readers, vertically rising rootobars (as in plot 7) would have great advantages. Otherwise, if the considerations to be discussed in the next section do not interfere, either type of rootobar plot (as in plot 3 <u>or</u> plot 7) would seem very desirable.

11D. Intervals of different lengths

There are situations where -- sometimes for extraneous reasons, sometimes because of the nature of the distribution, we want to use counts corresponding to varying intervals -- to intervals not all of the same length. Doing this raises some new problems -- and eventually throws a clearer light on what we are indeed doing.

If we are concerned with a histogram, there is little if any difficulty in using intervals of different lengths. We have only to recall that we are really using an area scale, so that the heights we plot have to be in a constant proportion to the ratios:

$$\frac{\text{number of values counted}}{\text{length of interval considered}}$$

Plot 8 shows how this works for our continuing example. In this plot we have used intervals for positive values just half the width of those used before, while continuing to use the old familiar intervals for negative values.

Except for an increased irregularity from rectangle to rectangle where the bases have been halved, the general impression given by this plot resembles reasonably closely that given by plot 5. This is clearly as we would hope it to be, though the increased irregularity can be the source of some difficulty.

Table 9 shows the details and the almost-trivial arithmetic that is required.

The calculations necessary to make this plot have caused us to label the vertical scale differently than in plot 5 (and plots 4, 6, and 7). We have been more careful and correct by saying that the scale is really a scale of "counts per unit base" instead of a scale of "counts", and by showing the length of the unit base to which the counts apply.

We probably should have done just this in all the earlier plots.

Making a bar chart for intervals of different length raises no issues of arithmetic, but can raise issues of policy. If, as we are now convinced, we are trying to plot a rate -- the rate at which we encounter observations as we move along the scale, to speak of "counts per unit base" essentially admits

plot 8 of chapter 11

Plot 8

The by now familiar example: histogram with intervals of varying length

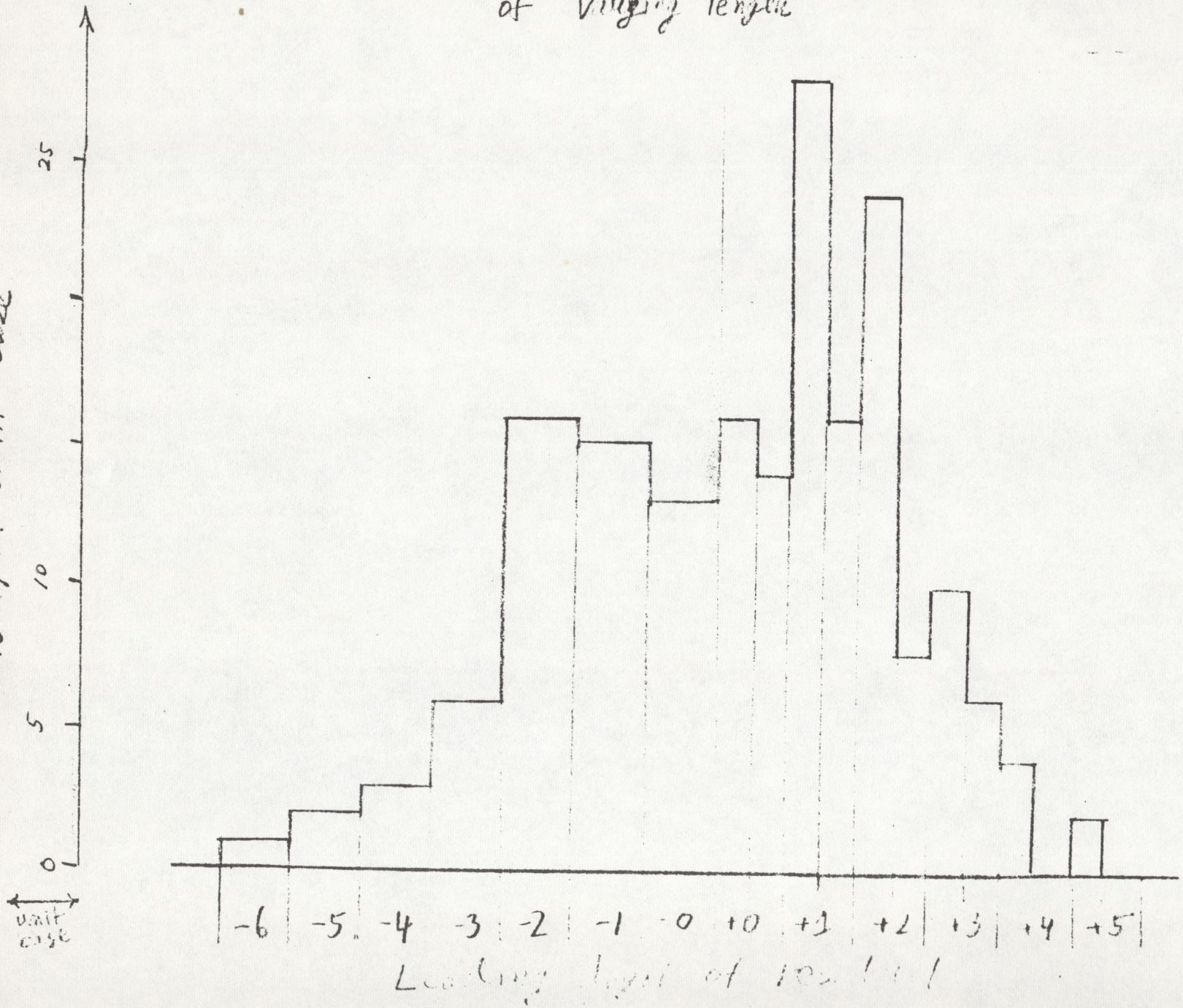


table 9 of chapter 11

Data and Arithmetic for the Same Old Example When Positive Intervals Are of Half Width. (Counts can be checked from Table 1.)

Leading digit	Next digits	Count of values	Counts length	(*)
+5	0 to 4	1	2	1.4
+4	5 to 9	0	0	0.0
+4	0 to 4	2	4	2.0
+3	5 to 9	3	6	2.4
+3	0 to 4	5	10	3.2
+2	5 to 9	4	8	2.8
+2	0 to 4	12	24	4.9
+1	5 to 9	8	16	4.0
+1	0 to 4	14	28	5.3
+0	5 to 9	7	14	3.8
+0	0 to 4	8	16	4.0
-0	all	13	13	3.6
-1	all	15	15	3.9
-2	all	16	16	4.0
-3	all	6	6	2.4
-4	all	3	3	1.7
-5	all	2	2	1.4
-6	all	1	1	1.0

(*) To be revealed in due course.

that this IS our interest when we use histograms -- we are supposed to copy the procedure for histograms as closely as we can. In particular, we can use the "counts/length" column of table 9 to set the height of a bar chart presentation of the same example. Plot **10** does just this.

What of rootograms? Again nothing new, once we admit that the vertical scale is:

THE SQUARE ROOT of COUNTS/LENGTH

The last column of table 9, marked (*), contains the square roots of the values in the previous column, all ready for use. Plot **11** shows the results of using them.

When we look carefully at the 2x2 array of plots of which this is the SE corner, namely:

	Equal intervals	Unequal intervals
Histogram:	plot 5	plot 8
Rootogram:	plot 6	plot 11

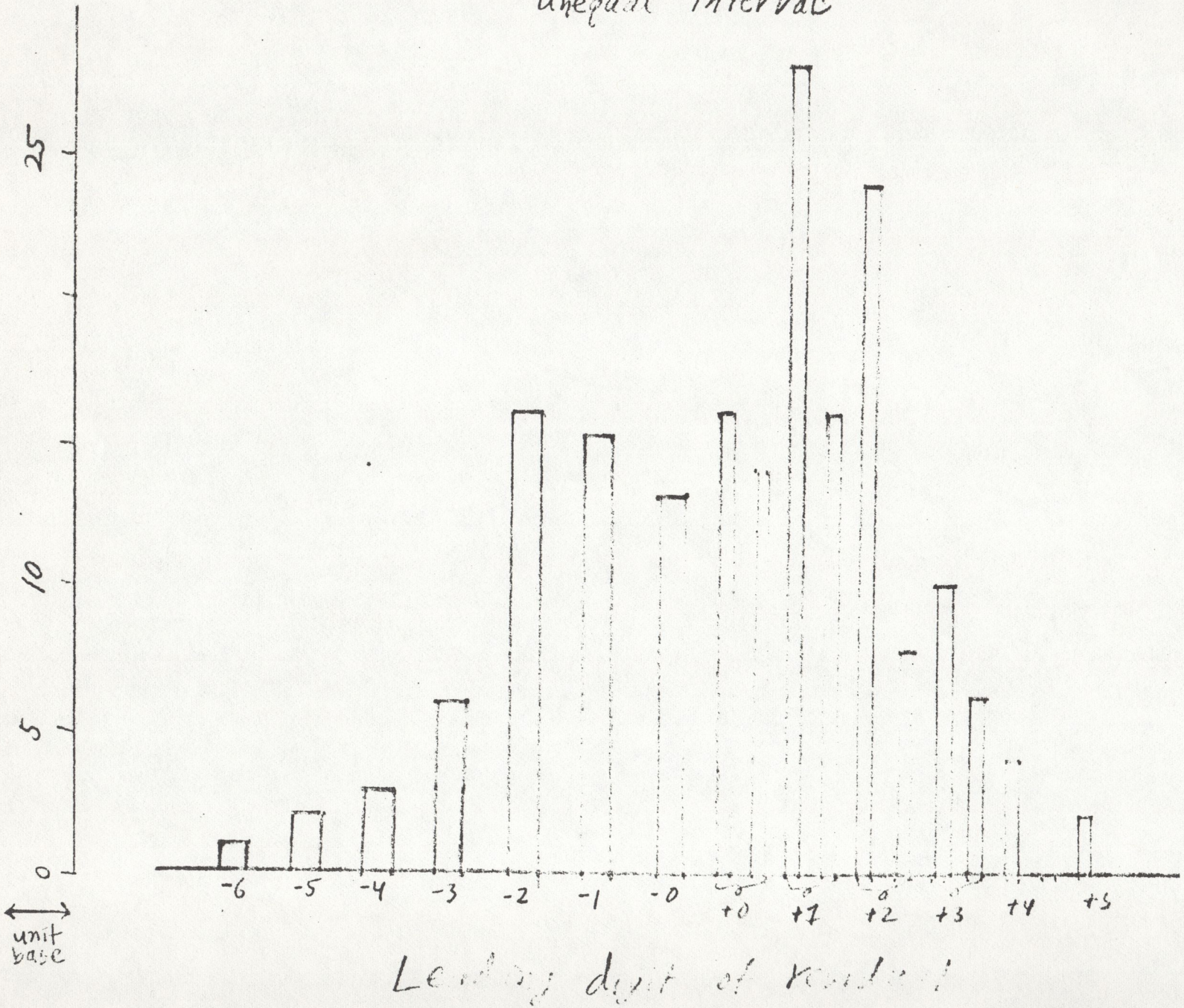
we realize an important fact:

- The effect on the rootogram of introducing unequal intervals is markedly less than the corresponding effect on the histogram.

Intervals of different widths lead, in all the techniques we are here considering, to different amounts of irregularity. For the histogram, this difference is combined with already existing differences in size of irregularity

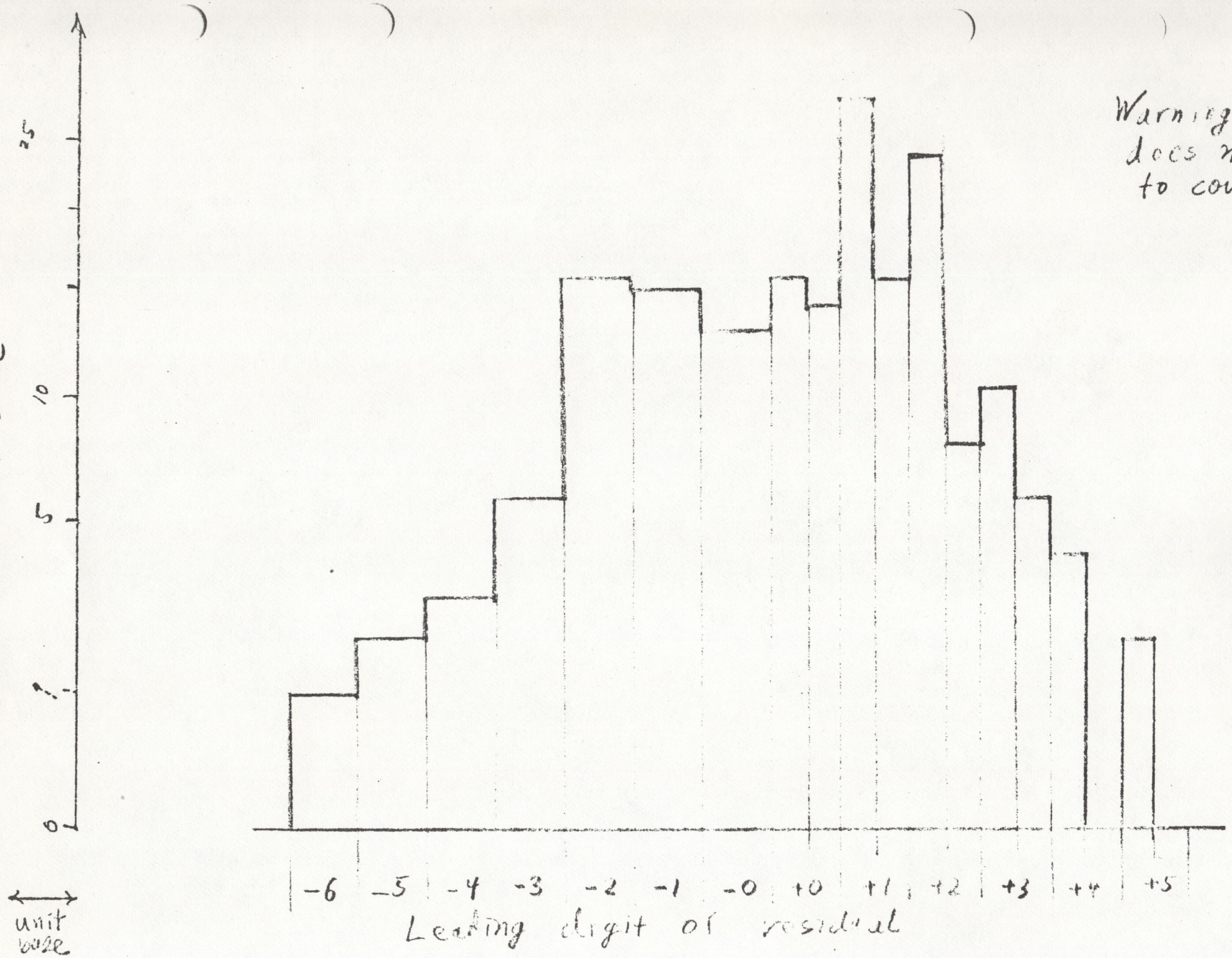
Plot 10

The by-now-familiar example: bar chart for unequal intervals



plot 10 of chapter 11

plot **11** of chapter 11



Plot 11

The by now-familiar example: histogram for unequal intervals

(larger where values are more frequent, smaller where they are not). Each source of difference in size tends to make the effects of the other worse. For the rootogram, which shows little change in size of irregularity when intervals are equal, the effects of inequality of interval width are less violent and less confusing. Again the rootogram proves more successful.

With both the square root and the counts/length ideas cleared up, a rootobar plot offers no difficulty. Plot **12** shows the result. This example displays our first sight of a general difficulty with rootobar plots. The total length of vertical lines in this picture is so great -- in comparison with the horizontal extent -- as to make the picture "busy" and somewhat confusing. After all, a roototbar plot requires about twice as many inches of vertical line as does the corresponding roototgram. We ought not to be surprised to find that we cannot afford to use as many intervals in a rootobar plot of given width and vertical scale as we can the corresponding rootogram.

Two ways lead out of such a difficulty:

- shrinking the vertical scale, thus reducing impact by making the picture too long for its height.
- decreasing the number of bars.

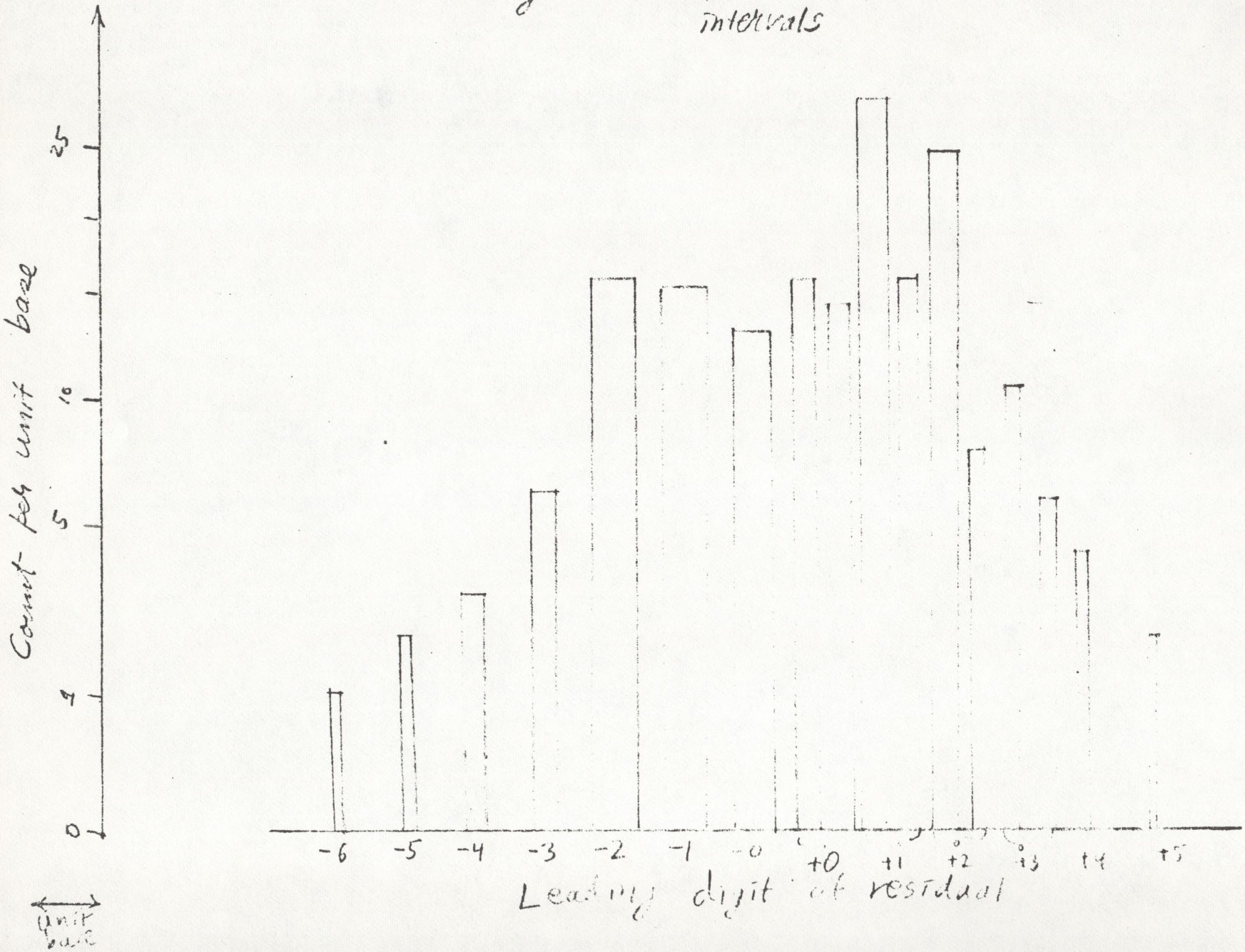
Sometimes we can be happy to follow one of these, sometimes not.

11E. <u>How to picture and compare distributions when the responses are ordered, not numbered</u>

Not all responses come as numbers -- and when they do, as we ought now to be keenly aware, they may not be the best numbers for analysis. If the responses are those of a

Plot 12

The by-now-familiar example: rootogram plot for unequal intervals



plot **12** of chapter 11

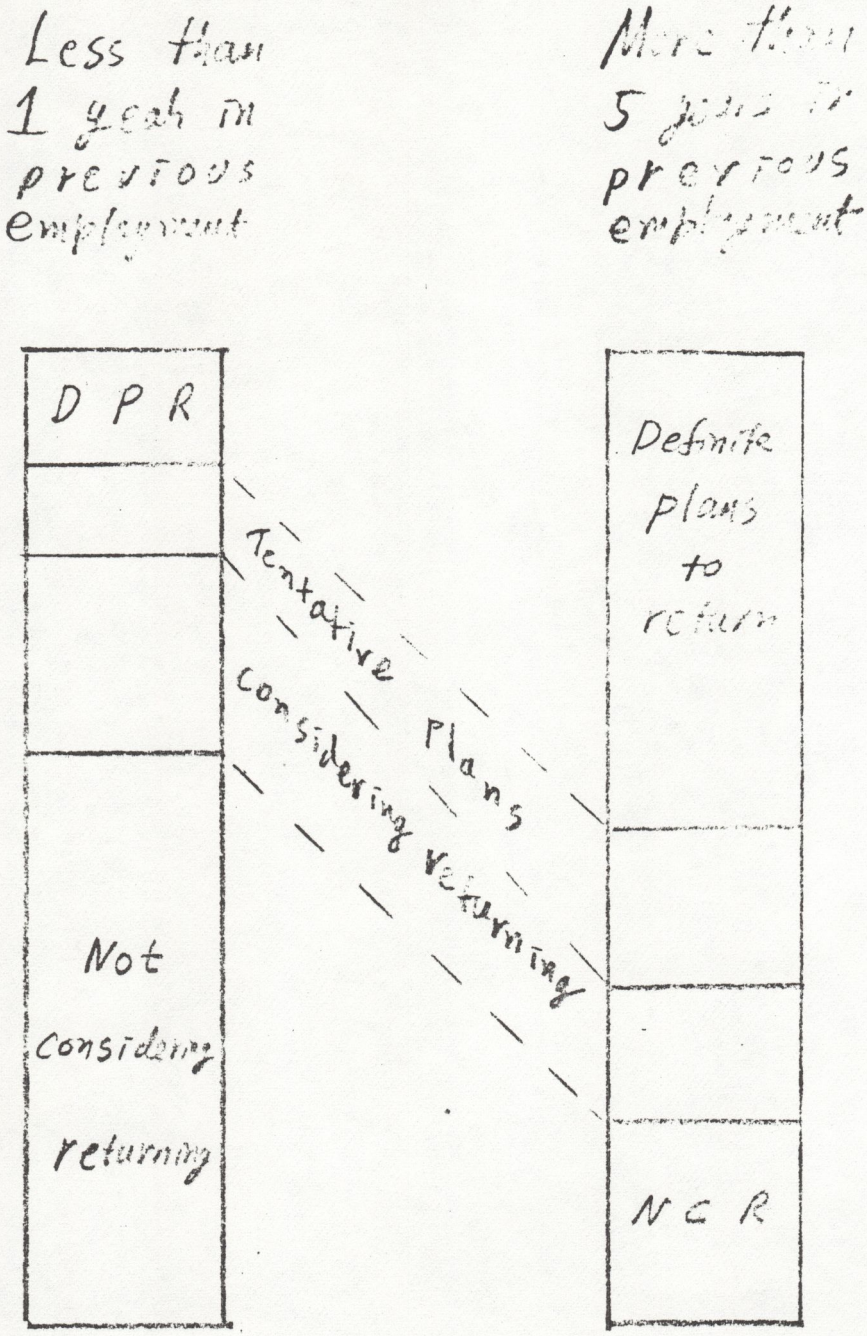
mixed fruit sorter: apple, orange, banana, pear, etc. we are not likely to do too well with a graphical approach. There does not seem to be any natural order for these kinds of fruit. When, however, there is a natural order among the possible responses, we can put it to good use graphically -- whether or not it comes from an initial set of numbers.

To do this, we have only to invert our manner of graphing. We have had the response on a fixed scale, and shown the frequency of occurrence by the areas of bars (or other geometric figures) placed according to response. With no numerical response -- either in fact or by desire to avoid any -- we cannot do this. We can, however, draw a fixed area of fixed shape to represent all the cases, and then show on these fixed areas where the cutting points fall.

If we are naive enough, we will use rectangular areas. Only a little thought is needed to see that the rectangles should go up and down, and be divided horizontally. Plot 13 shows how this could be done for the July separatees from the World War II army -- data already discussed from a different angle in section 10L. (Numbers in table 12 of chapter 10.)

By now we are not naive. We would like to plot something more sensible than % vertically. Yet we would like to retain the area principle. To do these both is not difficult -- we have only to give up rectangles, choosing instead some shape for which area is proportional to count while the vertical scale is more suitable than %.

Plans to return to previous employment of July separatees from World War II army.



We have already looked at this data -- and found that f-roots seem to describe it well (see section 10L). Accordingly, plot **14** shows the same information as plot 13, but with a vertical f-root scale and proportional area. Since only one % falls outside of 20-to-80% and that one not violently, we cannot be surprised to find rather little difference between the appearances and impacts of plots 13 and 14. More extreme examples can -- and do -- reveal more striking differences.

If we are to be able to make plots like plot 14 -- and analogous plots using f-log scales -- we need information about the variation of width with height that keeps area proportional to count. Once we have it, this information can be used for "split cigars" as in plot 14 (the pattern of choice for comparing two conditions), for "full cigars" (often helpful when more than two conditions are to be compared) or for any other shapes we may invent.

Table **15** provides this information.

11F. How to deal with distributions of counts -- especially when J-shaped

Distributions of counts do not always require special treatment. Data such as that of Brambell and Rowlands on litter size in bank voles - see problem 3 below, is a case in point. But distributions of counts frequently turn out to have one very, very long tail -- very large numbers of counts occurring very occasionally and very few counts occurring quite often. Indeed the name "J-shaped distribution" was invented to label such situations.

The same data as plot 13 more sensibly plotted.

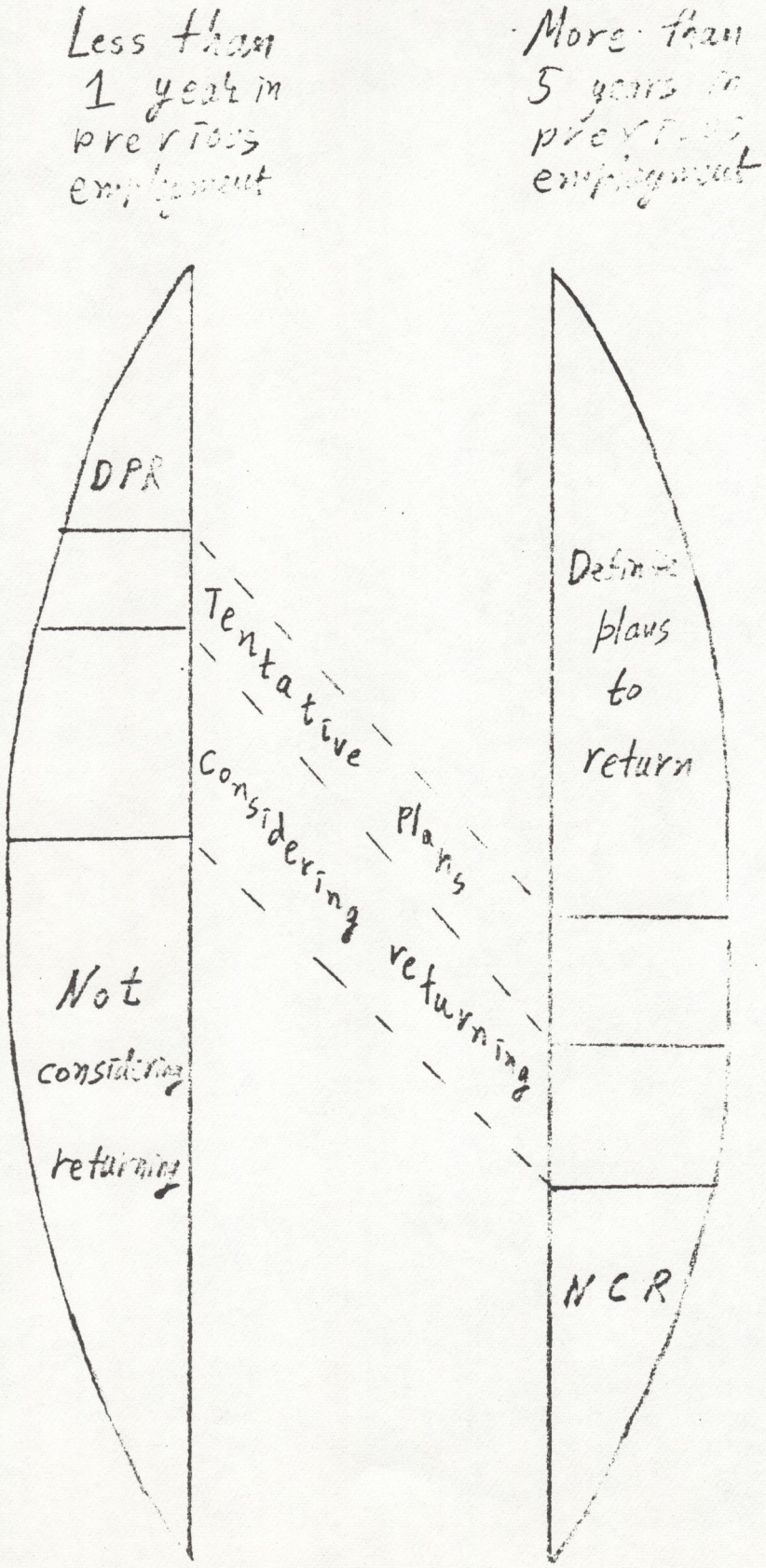


table 15 of chapter 11

Relative Widths at Various Heights that Keep Areas Proportional to Counts

Relative widths	Height: For f-roots	Height: For f-logs
1.00	±.00	±.00
.95	±.36	±.23
.90	±.50	±.33
.85	±.62	±.41
.80	±.71	±.48
.6	±.98	±.75
.4	±1.16	±1.03
.2	±1.31	±1.44
.1	±1.36	±1.82
.08	±1.37	±1.94
.06	±1.38	±2.08
.04	±1.39	±2.29
.02	±1.40	±2.64
.01	±1.41	±2.99
.005	±1.41	±3.34
.002	±1.41	±3.80
.001	±1.41	±4.20
.000	±1.41	(∞)

When one plots such a distribution in a conventional way all that can be seen is the long tail. Nearly all details of distribution -- and nearly all possibilities of effectively comparing different distributions -- are as if they had never been. Some better way to picture the situation is important.

Similar phenomena occur in connection with distributions of sizes of certain things. The sizes of cities and towns are counts, but, although the sizes of lakes, ponds, and other natural bodies of water are not counts, their distribution has a very similar shape. One long-tailed distributions of things that can only be positive -- or only positive and zero -- pose us a serious problem.

One approach is -- as always -- to take logs. The distribution of the log of a count -- or the log of a size -- is frequently well enough behaved for us to examine it -- or compare them -- effectively using conventional methods. When this both works and suffices, all is well.

In many cases, there appear to be finer details in the observed distributions than this approach appears to explain. (We are far from knowing what we should about what, in detail, we ought to expect from counts when their logs seem to be distributed in a well-behaved way.) Accordingly, at least until we learn more, there is good reason for another string to our bow, in the form of an alternate approach.

A toy example can make clear the change we are to consider. In 1951, Lumsden reported on the number of various kinds of mosquitoes caught at Tokwe, Bwamba, Uganda during a

five-day period (9th-14th July 1948). One catching station was described as "forest canopy" and at it species of Aedes were taken (in addition to various species of Taeniorhynchus and Culex). The numbers caught were:

A. longipalpis	13
A. africanus	7
A. nigerrimes	1
A. grahami	1
A. simpsoni	1
A. apicoargenteus	1

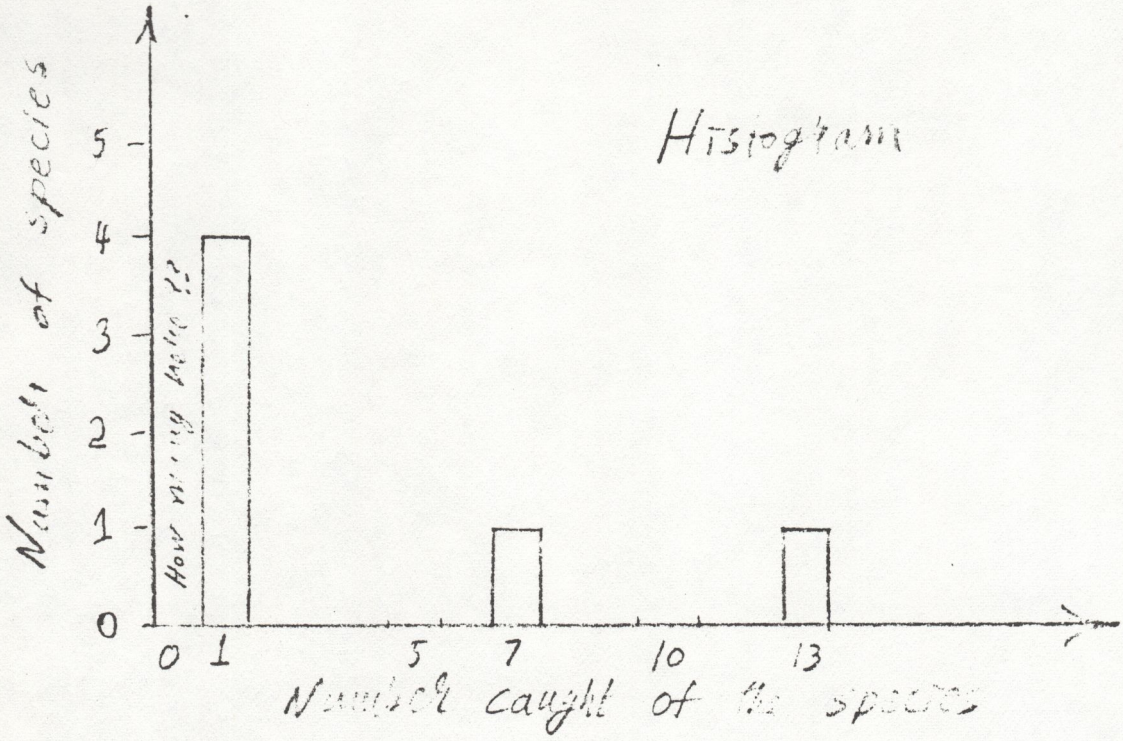
The usual histogram or rootobar plot would appear as in plot **16**. In it the focus of attention is the species. We find the area principle satisfied when, as in each picture of plot 16, each species is represented by the same area.

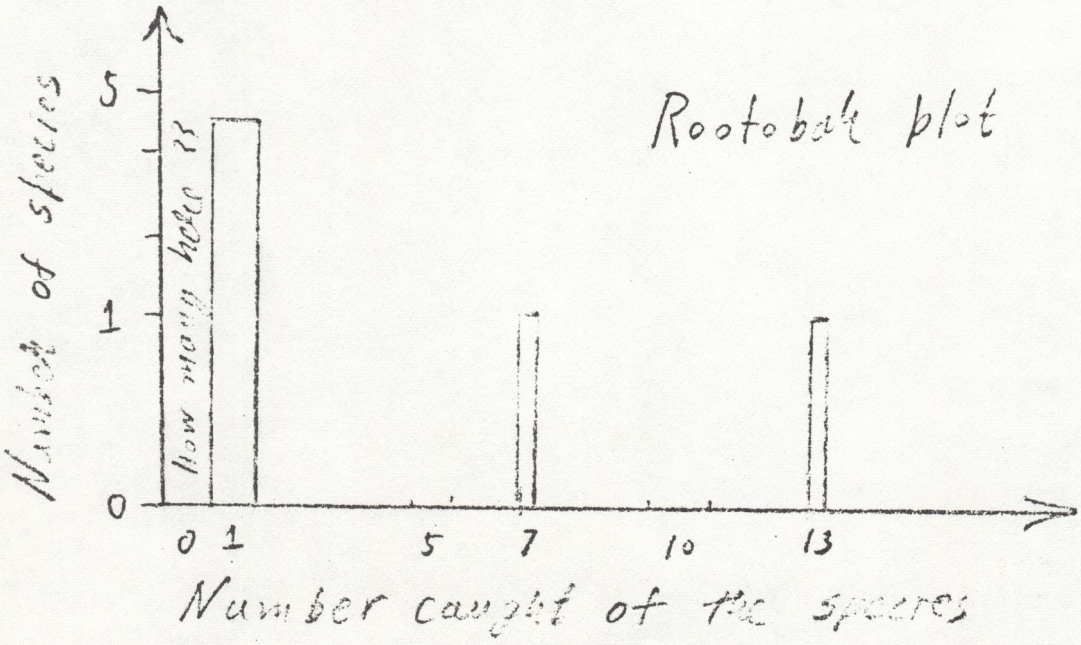
As part of taking a new approach, we can -- and will -- shift this focus, changing from one species to one mosquito. Now we shall find the area principle satisfied if each thing that is counted -- here each individual mosquito -- is represented by equal area. The naive form of approach to which this leads is one in which we number the "species" beginning with the one with the largest count and continuing in decreasing order (just as we found it natural to arrange the Aedes species above). This (inverse) rank we shall call here merely "rank".

The naive form of the new approach, then, is to plot "size" against "rank". Plot **17** does this for our toy example. Except for swapping what was close to the origin with what was far away, this change in approach has merely taken one

plot 16 of chapter 11

Conventional plots for catches of *Aedes* mosquitos

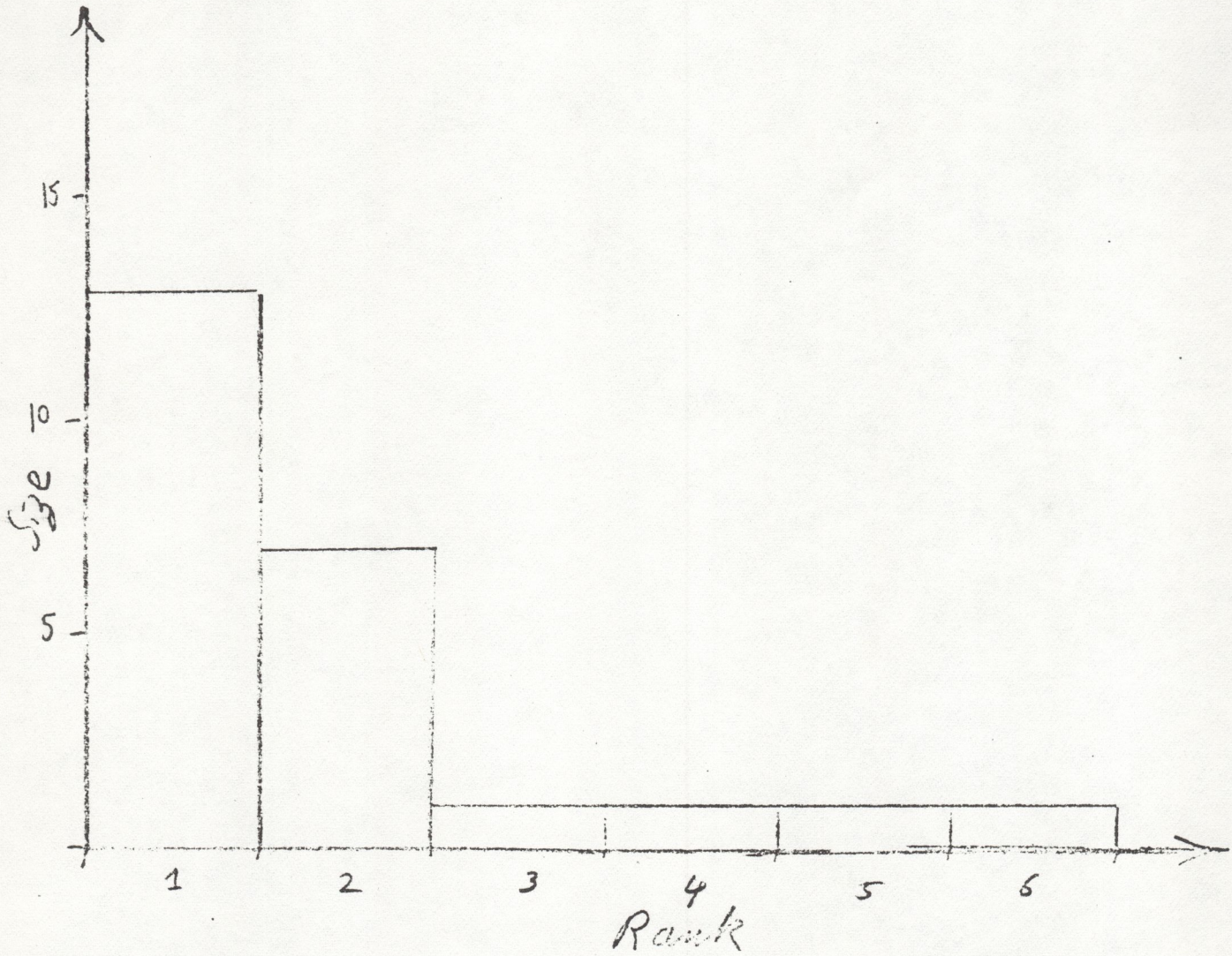




Source: W. H. R. Lumsden 1951. Probable insect vectors of yellow fever virus, from monkey to man, in Bwamba County, Uganda. Bull. Entomolog. Res. 42: 317-330. (Spec. first column of Table III at page 323.)

plot **17** of chapter 11

The new approach, plot of size vs. rank; area proportional to number of occurrences



J-shaped curve and replaced by another. If the first was very J-shaped, so too will be the second. How have we gained?

Our gain comes only when we note -- or recall -- that in such situations the product of rank and size often varies far less violently than either alone. In our toy example, we have the following:

Rank(s)	Size	Product
1	13	13
2	7	14
3 to 6	1	3 to 6

Clearly here rank-times-size varies by only a factor of 4, which is much better than the factor of 13 for raw counts.

Clearly we are now thinking of plotting rank-times-size vertically. If we are going to keep the area principle, we ought to arrange for the base of the rectangle associated with rank r to be of width proportional to 1/r. No one of us wants to go around adding up the reciprocals of the integers (in order to make a scale along which the rectangles could rub elbows). If we are prepared to make a reasonable approximation, no one of us will have to do this. It turns out that

$$\log(2r+1) - \log(2r-1)$$

is reasonably closely proportional to 1/r, as table 18 shows. Except for r = 1, the failure of proportionality is by much less than three parts in 100 and is thus quite undetectable graphically. (Those who worry about r = 1 can start the interval for r = 1 at any value between 1.08 and 1.10 to attain better than 3 in 100 precision everywhere.)

table 18 of chapter 11

Relationship Between log(2r+1) - log(2r-1) and 1/r

a) UNADJUSTED LOGS

r	log(2r+1) - log(2r-1)	1/r	Ratio
1	.47712	1	.4771
2	.22185	.5	.4437
3	.14613	.33333	.4384
4	.10914	.25	.4366
5	.08715	.2	.4357
6 to 10*	.28083	.64564	.4350
11 to 20*	.29056	.66877	.4345
(large)			(.4343)

* At these lines we show -- rather than individual values -- the sums of the values belonging in the second and third columns.

b) ADJUSTED LOGS (log 1.1 in place of log 1, log 3.02 in place of log 3)

1	.4383	1	.4383
2	.2192	.5	.4384
3	.14613	.33333	.4384

(and so on)

c) NOTE

logs above are to the base 10.

We thus preserve the area principle -- in its second form -- if we plot the rth largest "species" as a rectangle with dimensions:

- height proportional to r-times-size.
- base extending from log(2r-1) to log(2r+1).

If we now try to make a plot for our toy example, we are left with the situation shown in plot **19** where the product of rank-times-size for the four species tied at one individual each is shown as somewhere between 1-times-3 and 1-times-6. "What to do in such situations?" is a question best avoided in many circumstances.

We can avoid this neatly by plotting "halfway points" -- which occur whenever we come to a change in size (necessarily a decrease) -- and not plotting the values for which the problem arises. Thus if size 4 corresponds to ranks 637 to 792 and size 3 corresponds to ranks 793 to 1521, then, while it is hard to be sure about ranks for size 3 or size 4, "size 3.5" falls between ranks 792 and 793 and deserves "rank 792.5". Thus 3.5x792.5 would be a plausible "size x rank".

To simplify arithmetic it is easier to work with

$$(4+3)(792+793)$$

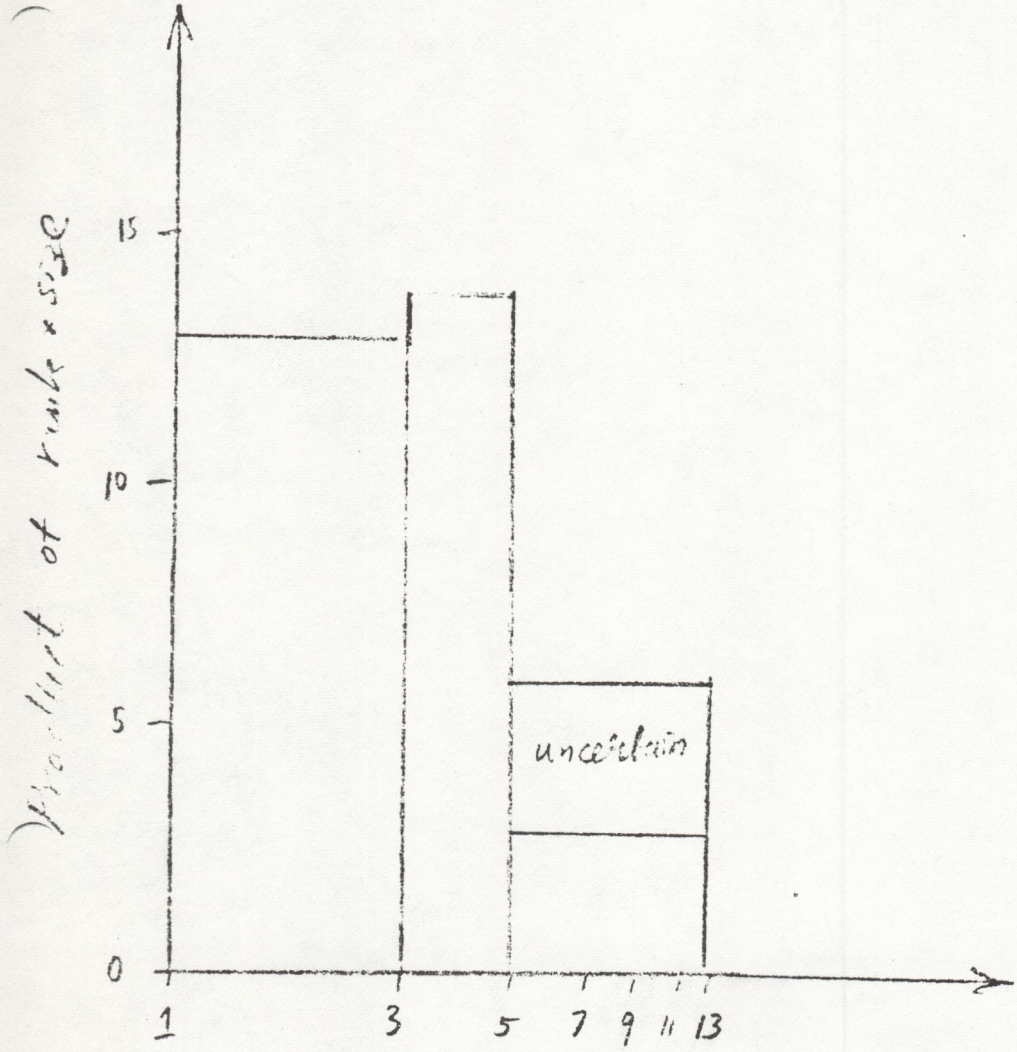
than with

$$\left(\frac{4+3}{2}\right)\left(\frac{792+793}{2}\right)$$

which is always exactly 1/4 of the arithmetically simpler expression.

plot **19** of chapter 11

Name plot of rank × size against the logarithms of odd integers (Trouble with ties)



odd integers (logarithmic scale)

Note: Rank r from log(2r-1) to log(2r+1).

If we plan to do this, we can supplement these "halfway points" with a few examples of well defined points associated with a size and a rank, usually for only the very smallest ranks. To make plotting compatible we double these up too, so that if the largest sizes were 4321 and 2314 we would plot

(4321+4321)(1+1) at log(1+1)

and

(2314+2314)(2+2) at log(2+2)

as well as

(4321+2314)(1+2) at log(1+2)

We can do this sort of thing:

- when there is only one instance of a size (so there is a unique rank).
- when there are exactly two instances of a size (so there are exactly two ranks, so that we should use their sum for "sum of ranks").

What we will want to make, then, is a (sum of sizes) x(sum of ranks) against log(sum of ranks) plot, something we shall abbreviate as a "rank-size-log-rank" plot. (The log scale for "sum of sizes" still begins at log 1 = 0 -- or, if you wish to be fussy, at $\log_{10} 1.1 = 0.041$ -- and ends at log(2n+1) where n species are represented by at least one individual.)

The picture for our toy example is now as shown in plot **20**, the calculations being shown in table **21**. The general impression given is of a plot with which one can distinguish different shapes of even very J-shaped distributions. Moreover, the area principle applies in its second form.

Let us then consider an example of more realistic size. We shall use a word-frequency example, leaving further discussion of the general subject to the next chapter (see section 12??).

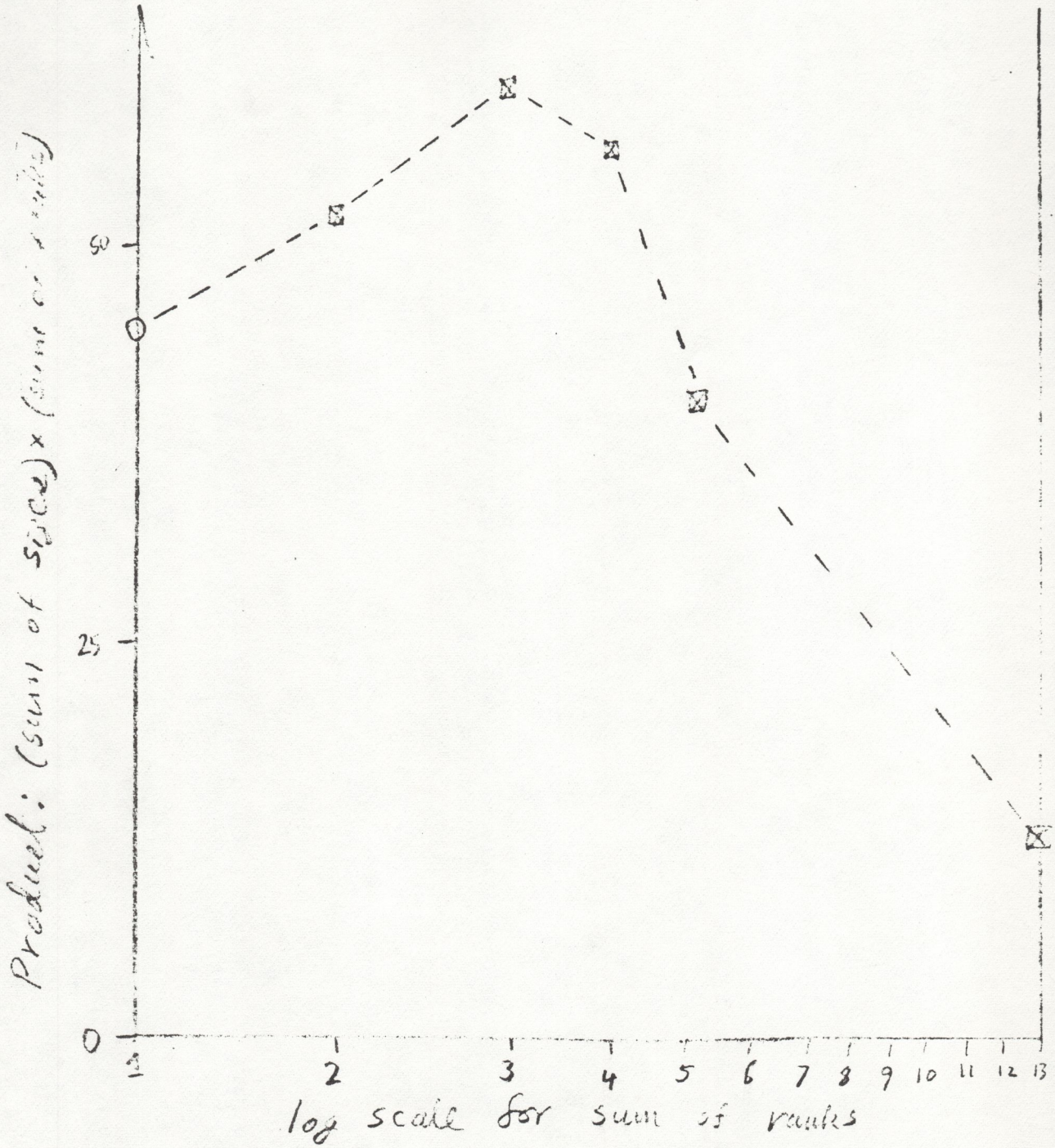
In 1953, Josselson (as reported by Herdan (1966), pages 96-97,174) counted the occurrence of the Russian words in Pushkin "The Captain's Daughter". Table **22** summarizes the results of the count and the calculation of some points for the rank-size-log rank plot. The resulting (selected) points are shown in plot **23**, where we see a marked constancy of the product between sum of ranks 25 and sum of ranks 3000. As is so often the case, however, the curve dips at both ends of the scale.

Since this plot is (nearly enough) area-true to occurrences, we can, for example, judge that about 1/2 of the area lies on each side of sum-of-ranks = 200. Thus the 100 most frequent words account for about 1/2 of all occurrences.

11G. <u>How far have we come</u>?

This chapter has developed a variety of pictorial approaches to the display of information about distributions when there is nothing with which we are to compare.

Rank-size-log rank plot for the Acton data (Final form)



⊠ Points from table 21

O Extrapolated point (see (b) of Table 31)

table **21** of chapter 11

Calculations for the Toy Example

Rank	Size	Sum of ranks	Sum of sizes	Product
(a) SUMS OF RANKS AND SIZES.				
1	13	2	26	52
		3	20	60
2	7	4	14	56
		5	8	40
3 to 6	1			
		13	1	13
7 on	0			

(b) STRAIGHT-LINE (in r) EXTRAPOLATION to r = 1, log r = 0.

Sum of ranks	Product
3	60
2	52
1	? (to be found)

$$? = 52 - \frac{60-52}{3-2}(2-1) = 44$$

table 22 of chapter 11

Summarization of Josselson's Frequency Counts for Pushkin's "The Captain's Daughter"

Size(s)*	Ranks	Sum of Sizes	Sum of Ranks	Product
			(1)	(3,487)
1160	1	2320	2	4,640
		1931	3	5,793
771	2	1542	4	6,164
		1495	5	7,475
724	3			
		1305	7	9,135
581	4			
		1060	9	9,540
479	5			
		909	11	9,999
430	6			
		851	13	11,063
421	7			
		800	15	12,000
379	8			
		676	17	11,492
297	9			
		586	19	11,134
289	10			
		577	21	12,111
288-282	11-12			
		539	25	13,475
257-244	13-15			
		480	31	14,880
236-187	16-20			
		365	41	14,965
178-119	21-30			
		237	61	14,457
118-90	31-40			
		179	81	14,499
89-73	41-50			
		145	101	14,645
72-62	51-60			
		122	121	14,762
60-49	61-70			
		97	141	13,677
48-43	71-81			
		84	163	13,692
41-40	82-90			
		78	181	13,766
38-37	91-100			
		73	201	14,673
36-35	101-111			
		69	221	15,249
34-33	112-117			
		65	241	15,665

Table 22 (Cont'd)

Size(s)*	Ranks	Sum of Sizes	Sum of Ranks	Product
32-31	118-134	61	269	16,409
30-29	135-149	57	299	16,801
28-27	150-158	53	317	16,801
26-25	159-169	49	339	16,611
24-23	170-179	45	359	16,155
22-21	180-196	41	393	16,113
20-19	197-216	37	433	16,021
18-17	217-242	33	485	16,005
16-15	243-271	29	543	15,747
14	272-293	27	587	15,849
13	294-313	25	627	15,675
12	314-346	23	693	15,939
11	347-380	21	761	15,981
10	381-419	19	839	15,941
9	420-460	17	921	15,651
8	461-539	15	1079	16,165
7	540-621	13	1243	16,159
6	122-735	11	1471	16,181
5	736-881	9	1763	15,867
4	882-1119	7	2239	15,673
3	1120-1552	5	3105	15,525
2	1553-2399	3	4799	14,397
1	2400-4783	1	9567	9,567

* Only sizes (= numbers of occurrences) actually found are mentioned.

** Correcting an obvious misprint.

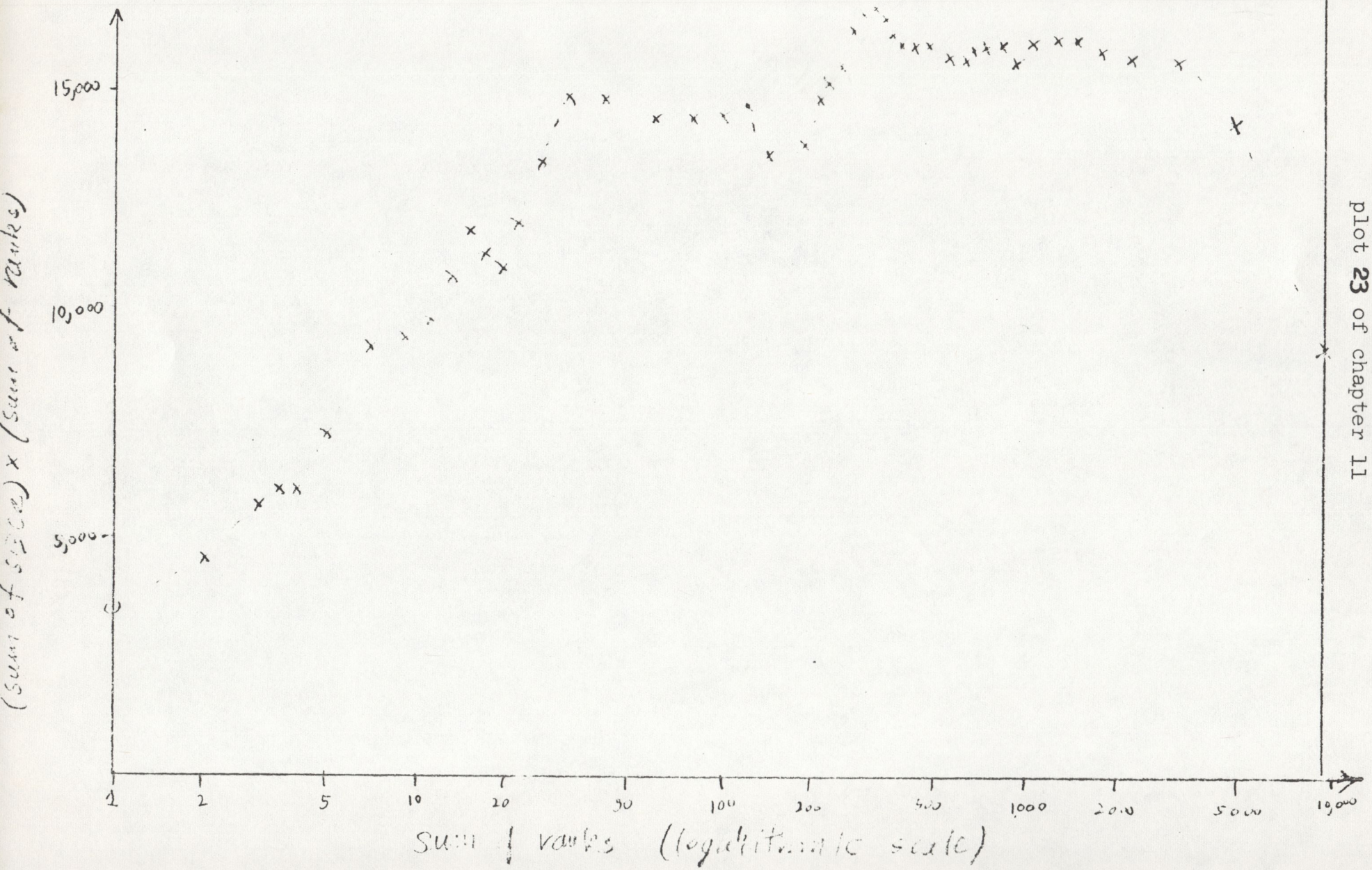
Source: G. Herdan 1966. The advanced theory of language as choice and chance. (Kommunikation und Kybernetik in einzeldarstellungen. Band 4). Springer, New York. Spec. Table 32 at page 97.

plot 23 of chapter 11

Plot 23

Rank-size-log rank plot for occurences of words in Pushkin's

The Captain's Daughter



Three partially conflicting forces have to be resolved, often differently in different cases. They are:

- the usually overriding advantages of a square-root-of-counts-per-unit-length scale.
- the desirability of maintaining the area principle.
- the tendency of rootobar plots to become busy and confusing (and to involve more drafting effort).

If it were not for the third of these, we might always use rootobar plots. If it were not for the second, we might always use rootograms. If it were not for the first, we might always use histograms. In practice, situations that really indicate histograms are not too common -- the choice is usually between a rootobar plot and a rootogram. (Every analyst has to develop his own balance here.)

When any of these techniques are used to plot J-shaped distributions -- distributions with one long tail -- we are unlikely to see anything more than the distribution is J-shaped. Exploration of such distributions calls for a new kind of plot. We can do quite well by plotting "sum of ranks times sum of sizes" against the logarithm of the sum of ranks. We do this for all cases where the sum of two ranks, equal or adjacent, is well determined.

*****

11R. Xmas-tree patterns

In section 11B we went from stem-and-leaf displays to various forms of "bars-flying-from-a-mast", the latter being a step on the way to the usual bar plots. We could have gone to more symmetric displays along the way, but did not because such displays do not seem specially useful. They are used, however, and are thus worth passing attention.

Table 24 shows what happens when we try to make the stem-and-leaf pattern more symmetrical. Many displays in which left-to-right spread is symmetric about a vertical axis, of which this table illustrates only one kind, are likely to be called "Xmas trees" or "tree diagrams".

One important comment about the difference in character between table 1 and table 24 is in order. Table 1 is a worksheet -- possibly polished up a trifle -- that also serves as a graphic display. Table 24 is a picture -- it could not have been started until all 120 values are at hand, since neither end of each line is fixed until we know how many entries go on that line.

Plot 25 sets out, in terms of bars of constant width, the same overall impressions conveyed by table 24. Its pros and cons, in comparison with that table are mostly associated with choices of the ratio of width to height, choices that can clearly be made more freely and flexibly in a plot. Demographers like to use somewhat similar sorts of plot to show age distributions for the populations of various countries.

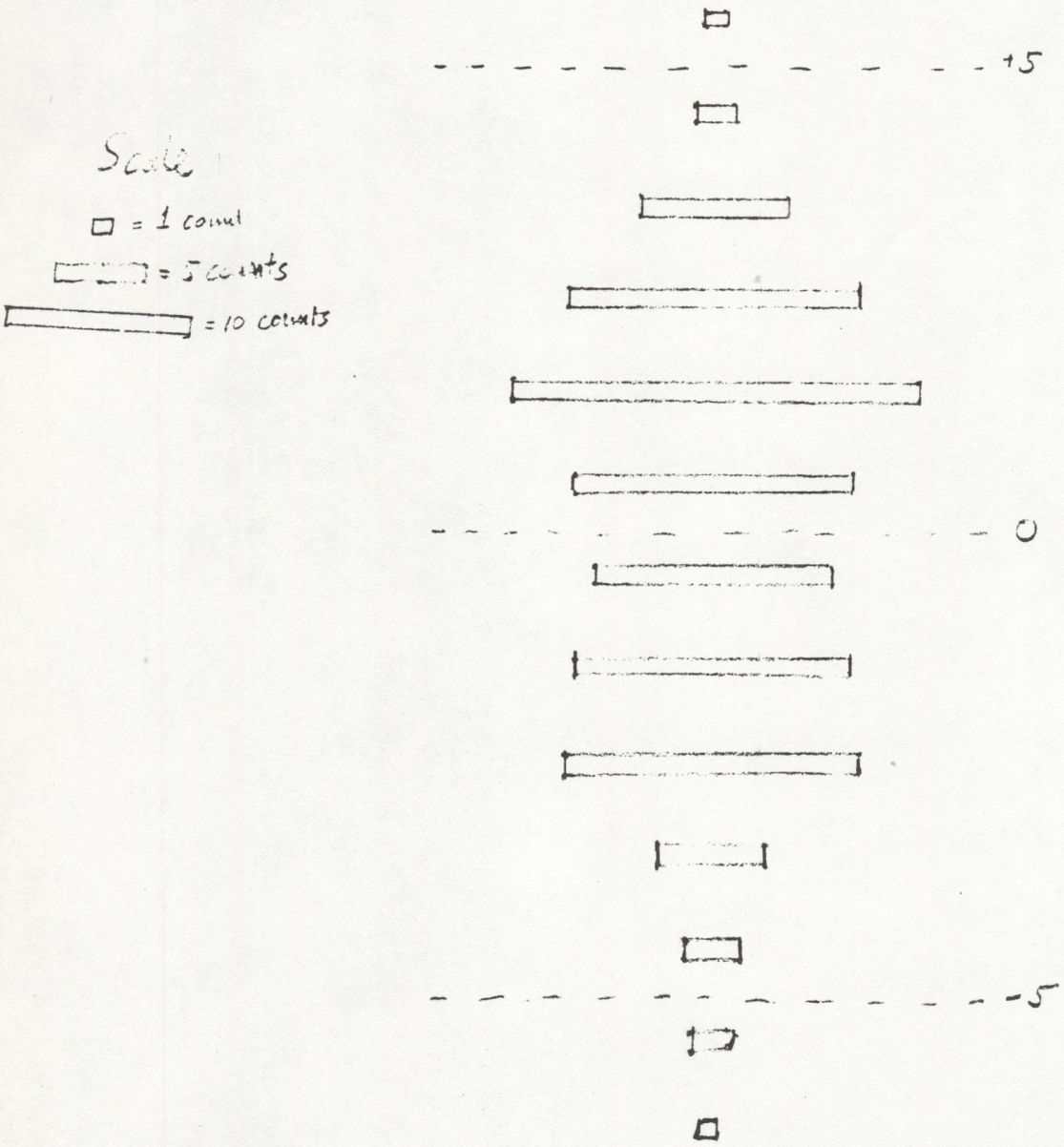
There is, of course, no reason why we should not use rootobars in a Xmas-tree display. Plot 26 does just this.

table **24** of chapter 11

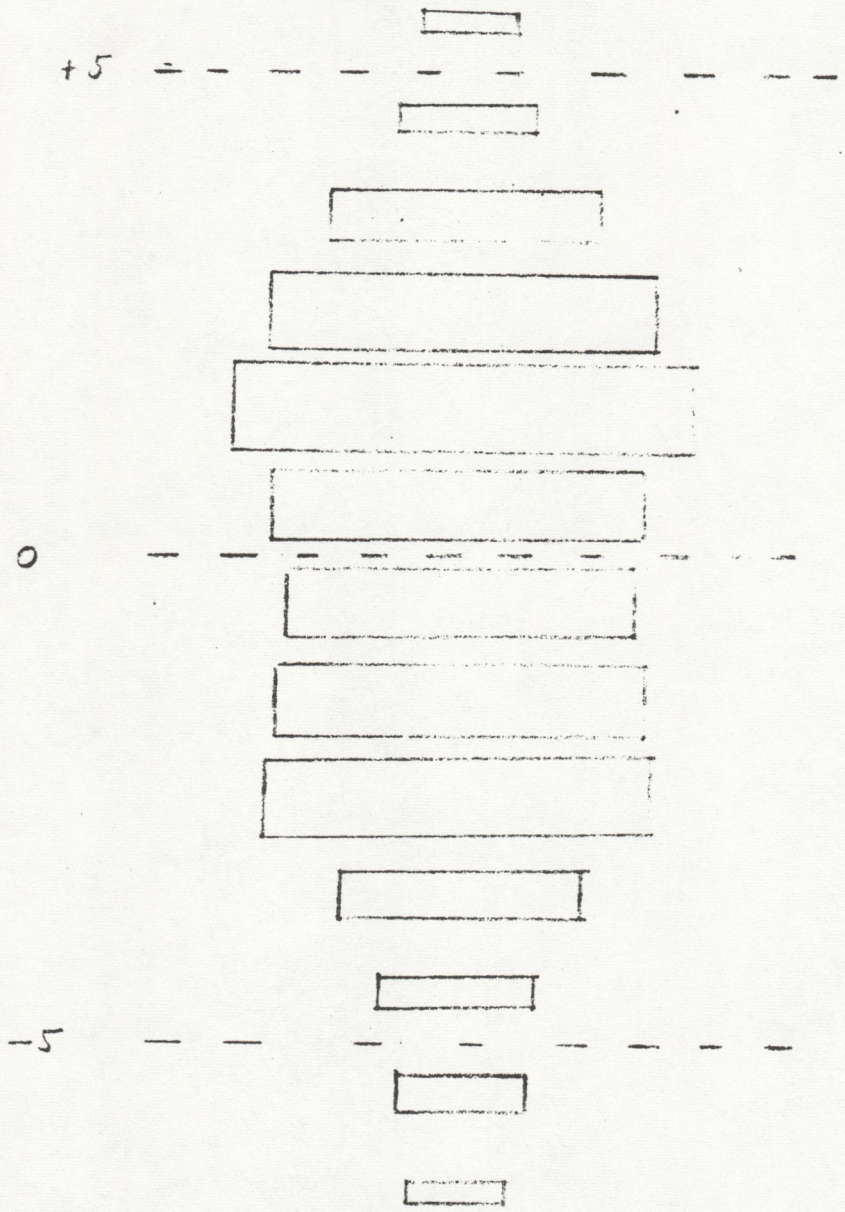
Symmetrical (Xmas-tree) Version of the Stem-and-leaf Array of Table 4 from Chapter 4
(The ... mark, from top to bottom, the heights corresponding to
+5, 0, and -5, 0 respectively)

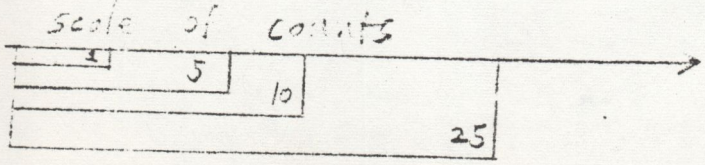
```
                4
               . . .
                2 2
            1 1 8 3 3 9 2 6
        7 0 2 4 3 1 4 4 8 6 3 3 8 4 2 0
    2 6 3 1 3 3 3 6 7 9 4 9 7 9 2 2 5 0 0 1 4 3
        1 8 1 4 1 3 6 2 4 2 6 7 7 6 8
                  . . .
          1 6 7 0 4 8 4 4 3 7 4 3 5
        2 6 6 1 2 8 3 0 6 5 7 6 6 1 6
        9 3 0 0 1 6 0 7 0 7 9 1 6 4 5 6
                4 5 2 2 1 2
                  2 0 3
                  1 7
                 . . .
                   1
```

Same distribution, summarized by bars of constant width.



The same data represented by rootbulls.
(Bulls each of whose dimensions is proportional to the square root of the corresponding counts)





*****

11Y. Problems

1. In 1936, Williams gave the following total numbers of Noctuidae (a category of moth) captured in each of eight periods of the night for: (i) all nights, (ii) all nights in the week before full moon, (iii) all nights in the weeks after full moon:

period:	1	2	3	4	5	6	7	8
All:	599	826	1813	1454	1759	1429	955	309
pre-full:	115	174	175	362	467	380	230	52
pre-new:	75	136	150	234	190	152	70	35

Display these counts so as to effectively reveal any change in distribution through the night between pre-full and pre-new weeks. Discuss. What other types of display might have been useful? Source: Williams, C. B. 1936. The influence of moonlight on the activity of certain nocturnal insects, particularly of the family Noctuidae, as indicated by a light trap. Phil. Trans. Roy. Soc. Lond. 226: 357-389, specially his table VII at page 374.

2. In 1940, C. I. Smith published the first of a series of papers on the Great Bahama Bank. In particular, he gave the number of days in each month during 1939 for which the wind at Nassau came from specified directions. His counts are given as table 23. Omitting days of calm, plot the counts as revealingly as you can. Why did you select this way of presenting the data?

table 23 of chapter 12

Wind Direction at Nassau, Bahamas Month-by-month During 1939

Month	N	NE	E	SE	S	SW	W	NW	(Calm)
Jan	3	6	5	7	1	1	0	2	(6)
Feb	1	7	4	12	1	2	0	0	(1)
Mar	1	9	7	10	2	1	0	0	(1)
Apr	1	5	5	11	3	3	1	1	(0)
May	3	3	6	10	3	5	0	0	(1)
June	0	4	8	10	2	1	1	1	(3)
July	0	0	5	14	5	5	0	0	(2)
Aug	0	3	7	12	3	4	0	0	(2)
Sept	1	7	2	7	2	1	0	0	(10)
Oct	0	21	4	1	2	0	0	0	(3)
Nov	6	12	3	3	1	0	1	4	(1)
Dec	5	5	5	5	2	2	0	7	(0)

From: C. I. Smith 1951. The great Bahama bank. I. General hydrographical and chemical features. J. Marine Res. 3: 147-170, specially his table III, at page 155.

3. Brambell and Rowlands give counts of the numbers of certain rodents -- bank voles (Evotomys Glareolus) -- according to litter size as estimated in three ways: by the number of corpora lutea, by the number of placental sites, and by the number of embryos found (both in their work and by earlier workers), namely:

No:	1	2	3	4	5	6	7	8	more	(Tot)
Corp. Lut.	1	8	43	108	82	22	7	2	(*)	(273)
Pla. sites	-	1	11	20	13	11	2			(58)
Embryos	1	1	16	27	21	4	(own work)			(70)
Embryos	1	4	29	35	14	5	(earlier)			

(*) 2 at 9, 1 at 11, 1 at 12.

Plot the distributions in a revealing way.

Source: Brambell, F. W. Rogers and I. W. Rowlands 1936. Reproduction of the bank vole (Evotomys Glareolus, Schreber. I - The oestrous cycle of the female. Phil. Trans. Roy. Soc. London 226: 71-98, specially their Table IV on page 79.

4. Display the distribution of logs of number of occurrences based on the data of table 40 of chapter 11 about the frequency of occurrences of words in The Captain's Daughter. (Further judicious summarization is appropriate.)

5. Wake (1957) gave frequencies of sentence lengths of two editions of Aristotle's Categories, as repeated in table 24. Make comparative rootograms of these distributions. Comment on the results.

table 24 of chapter 11

Distributions of Sentence Lengths for Two Editions of Aristotle's Categories

Sentence length (words)	Number of sentences Oxford	Loeb
4	2	3
5	7	12
6	15	15
7	21	20
8-10*	41	51
10*-12	35	49
13-15	52	51
16-19	48	42
20-25	55	41
26-31	23	26
32-39	11	12
40-50	1	1
51-63	1	1
64-79	-	2

Source: Wake 1957. Specifically Table 2 at page 336.